

# ALAGAPPA UNIVERSITY



[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category–I University by MHRD-UGC]

(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003

## Directorate of Distance Education

## M.Sc. (Mathematics)

### IV - Semester

### 311 43

# NUMERICAL ANALYSIS

**Authors:**
**Dr. N. Dutta,** Professor of Mathematics, Head - Department of Basic Sciences & Humanities, Heritage Institute of Technology, Kolkata
Units (2, 4, 6-8, 10-13)
**Vikas® Publishing House:** Units (1, 3, 5, 9, 14)






Vikas® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.
E-28, Sector-8, Noida - 201301 (UP)
Phone: 0120-4078900 • Fax: 0120-4078999
Regd. Office: A-27, 2nd Floor, Mohan Co-operative Industrial Estate, New Delhi 1100 44
• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

**Work Order No. AU/DDE/DE 12-02/Preparation and Printing of Course Materials/2020 Dated 30.01.2020 Copies - 1000**

# SYLLABI-BOOK MAPPING TABLE

## Numerical Analysis

| Syllabi | Mapping in Book |
|---|---|
| **BLOCK - I : POLYNOMIAL EQUATIONS AND EIGEN VALUE PROBLEMS**<br>**UNIT - 1**<br>Transcendental and Polynomial Equations: Rate of Convergence of Iterative Methods.<br>**UNIT - 2**<br>Methods for Finding Complex Roots - Polynomial Equations.<br>**UNIT - 3**<br>Birge - Vieta Method, Bairstow's Method, Graeffe's Root Squaring Method.<br>**UNIT - 4**<br>System of Linear Algebraic Equations and Eigen Value Problems: Error Analysis of Direct and Iteration Methods. | **Unit 1:** Transcendental and Polynomial Equations **(Pages 3-29);**<br>**Unit 2:** Methods for Finding Complex Roots and Polynomial Equations **(Pages 30-53);**<br>**Unit 3:** Birge – Vieta, Bairstow's and Graeffe's Root Squaring Methods **(Pages 54-65);**<br>**Unit 4:** Solution of Simultaneous Linear Equation **(Pages 66-85);** |
| **BLOCK - II : EIGEN VECTORS, INTERPOLATION, APPROXIMATION, DIFFERENTIATION AND INTEGRATION**<br>**UNIT - 5**<br>Finding Eigen Values and Eigen Vectors - Jacobi and Power Methods.<br>**UNIT - 6**<br>Interpolation and Approximation: Hermite Interpolations - Piecewise and Spline Interpolation - Bivariate Interpolation.<br>**UNIT - 7**<br>Approximation - Least Square Approximation and Best Approximations.<br>**UNIT - 8**<br>Differentiation and Integration: Numerical Differentiation - Optimum Choice of Step - Length - Extrapolation Methods. | **Unit 5:** Eigen Values and Eigen Vectors **(Pages 86-106);**<br>**Unit 6:** Interpolation and Approximation **(Pages 107-146);**<br>**Unit 7:** Approximation **(Pages 147-171);**<br>**Unit 8:** Numerical Integration and Numerical Differentiation **(Pages 172-220)** |
| **BLOCK - III : PDE, ODE AND EULER METHODS**<br>**UNIT - 9**<br>Partial Differentiation - Methods Based on Undetermined Coefficient - Gauss Methods.<br>**UNIT - 10**<br>Ordinary Differential Equations: Local Truncation Error - Problems.<br>**UNIT - 11**<br>Euler, Backward Euler, Midpoint, -Problems. | **Unit 9:** Partial Differential Equations **(Pages 221-283);**<br>**Unit 10:** Ordinary Differential Equations **(Pages 284-299);**<br>**Unit 11:** Euler's Method **(Pages 300-307)** |
| **BLOCK - IV: TAYLOR'S METHOD, R.K METHOD AND STABILITY ANALYSIS**<br>**UNIT - 12**<br>Taylor's Method -Related Problems.<br>**UNIT - 13**<br>Second Order Runge Kutta Method - Stability Analysis.<br>**UNIT - 14**<br>Stability Analysis. | **Unit 12:** Taylor's Method **(Pages 308-312);**<br>**Unit 13:** Runge Kutta Method **(Pages 313-321);**<br>**Unit 14:** Stability Analysis **(Pages 322-328)** |

# CONTENTS

# INTRODUCTION



Numerical analysis is the study of algorithms to find solutions for problems of continuous mathematics. It helps in obtaining approximate solutions while maintaining reasonable bounds on errors. Although numerical analysis has applications in all fields of engineering and the physical sciences, yet in the 21st century life sciences and both the arts have adopted elements of scientific computations. Ordinary differential equations are used for calculating the movement of heavenly bodies, i.e., planets, stars and galaxies. Besides, it evaluates optimization occurring in portfolio management and also computes stochastic differential equations to solve problems related to medicine and biology. Airlines use sophisticated optimization algorithms to finalize ticket prices, airplane and crew assignments and fuel needs. Insurance companies too use numerical programs for actuarial analysis. The basic aim of numerical analysis is to design and analyse techniques to compute approximate and accurate solutions to unique problems.

In numerical analysis, two methods are involved, namely direct and iterative methods. Direct methods compute the solution to a problem in a finite number of steps whereas iterative methods start from an initial guess to form successive approximations that converge to the exact solution only in the limit. Iterative methods are more common than direct methods in numerical analysis. The study of errors is an important part of numerical analysis. There are different methods to detect and fix errors that occur in the solution of any problem. Round-off errors occur because it is not possible to represent all real numbers exactly on a machine with finite memory. Truncation errors are assigned when an iterative method is terminated or a mathematical procedure is approximated and the approximate solution differs from the exact solution.

This book, *Numerical Analysis*, is divided into four blocks that are further divided into fourteen units which will help you understand how to solve transcendental and polynomial equations, rate of convergence of iterative methods, methods for finding complex roots – polynomial equations, Birge-Vieta method, Bairstow's method, Graeffe's root squaring method, system of linear algebraic equations and eigenvalue problems, error analysis of direct and iteration methods, finding eigenvalues and eigenvectors – Jacobi and power methods, interpolation and approximation, Hermite interpolations, piecewise and spline interpolation, approximation, least square approximation and best approximations, differentiation and integration, numerical differentiation, partial differentiation, ordinary differential equations, Euler, backward Euler, Taylor's method, second order Runge Kutta methods, and stability analysis.

The book follows the Self-Instruction Mode or the SIM format wherein each unit begins with an 'Introduction' to the topic followed by an outline of the 'Objectives'. The content is presented in a simple, organized and comprehensive form interspersed with 'Check Your Progress' questions and answers for better understanding of the topics covered. A list of 'Key Words' along with a 'Summary' and a set of 'Self Assessment Questions and Exercises' is provided at the end of the each unit for effective recapitulation. Logically arranged topics, relevant solved examples and illustrations have been included for better understanding of the topics.

NOTES

# BLOCK - I
# POLYNOMIAL EQUATIONS AND EIGEN VALUE PROBLEMS

# UNIT 1 TRANSCENDENTAL AND POLYNOMIAL EQUATIONS

**Structure**



## 1.0 INTRODUCTION

In mathematics, a **polynomial** is an expression consisting of variables (also called indeterminate) and coefficients, that involves only the operations of addition, subtraction, multiplication, and non-negative integer exponents of variables. An example of a polynomial of a single indeterminate, $x$, is $x^2 - 4x + 7$. An example in three variables is $x^3 + 2xyz^2 - yz + 1$.

Polynomials appear in many areas of mathematics and science. For example, they are used to form polynomial equations, which encode a wide range of problems, from elementary word problems to complicated scientific problems; they are used to define **polynomial functions**, which appear in settings ranging from basic chemistry and physics to economics and social science; they are used in calculus and numerical analysis to approximate other functions. In advanced mathematics, polynomials are used to construct polynomial rings and algebraic varieties, central concepts in algebra and algebraic geometry.

In this unit, you will study about transcendental and polynomial equations, and rate of convergence of iterative methods.

## 1.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand linear integral equations and some basic identities

NOTES

- Reduce initial value problems to Volterra integral equations
- Know the methods of successive approximation and successive substitution to solve Volterra equations of second kind, iterated kernels and Neumann series for Volterra equations
- Express resolvent kernel as a series in $\lambda$
- Know Laplace transform method for a difference kernel
- Find the solution of a Volterra equation of the first kind
- Reduce boundary value problems to Fredholm integral equations
- Know the method of successive approximation and successive substitution to solve Fredholm equations of the second kind
- Know iterated kernels and Neumann series for Fredholm equations
- Express resolvent kernel as a sum of series and Fredholm resolvent kernel as a ratio of two series
- Know Fredholm equations with separable kernels, approximation of a kernel by a separable kernel and Fredholm alternative

## 1.2 TRANSCENDENTAL AND POLYNOMIAL EQUATIONS

In mathematics, an integral equation is an equation in which an unknown function appears under an integral sign.

An integral equation in $u(x)$ is given by,

$$u(x) = f(x) + \lambda \int_{\alpha(x)}^{\beta(x)} K(x,t)u(t)dt \qquad \ldots(1.1)$$

where $K(x, t)$ is called the kernel of the integral Equation (1.1) and $\alpha(x)$ and $\beta(x)$ are the limits of integration. It can be easily observed that the unknown function $u(x)$ appears under the integral sign. It is to be noted here that both the kernel $K(x, t)$ and the function $f(x)$ in Equation (1.1) are given functions; and $\lambda$ is a constant parameter. We have to determine the unknown function $u(x)$ that will satisfy Equation (1.1).

An integral equation can be classified as a linear or nonlinear integral equation. The most frequently used integral equations fall under two major classes, namely Volterra and Fredholm integral equations. In this unit we will distinguish following integral equations:

- Volterra integral equations
- Fredholm integral equations

NOTES

### Volterra Integral Equations

The most standard form of Volterra linear integral equations is of the form

$$\phi(x)u(x) = f(x) + \lambda \int_a^x K(x,t)u(t)dt$$

where the limits of integration are function of $x$ and the unknown function $u(x)$ appears linearly under the integral sign.

If the function $\varphi(x) = 1$, then equation becomes

$$u(x) = f(x) + \lambda \int_a^x K(x,t)u(t)dt$$

and this equation is known as the Volterra integral equation of the second kind; whereas if $\varphi(x) = 0$, then the equation becomes

$$f(x) + \lambda \int_a^x K(x,t)u(t)dt = 0$$

which is known as the Volterra equation of the first kind.

### Fredholm Integral Equations

The most standard form of the Fredholm linear integral equations is given by,

$$\phi(x)u(x) = f(x) + \lambda \int_a^b K(x,t)u(t)dt \qquad ...(1.2)$$

where the limits of integration $a$ and $b$ are constants and the unknown function $u(x)$ appears linearly under the integral sign. If the function $\varphi(x) = 1$, then Equation (1.2) becomes,

$$u(x) = f(x) + \lambda \int_a^b K(x,t)u(t)dt$$

and this equation is called Fredholm integral equation of second kind; whereas if $\varphi(x) = 0$, then Equation (1.2) gives,

$$f(x) + \lambda \int_a^b K(x,t)u(t)dt = 0$$

which is called Fredholm integral equation of the first kind.

### Initial Value Problems Reduced to Volterra Integral Equations

Consider the integral equation,

$$y(t) = \int_0^t f(t)dt$$

**NOTES**

The Laplace transform of $f(t)$ is defined as

$$L\{f(t)\} = \int_0^\infty e^{-st} f(t)dt = F(s).$$

Using this definition the above integral equation can be transformed to

$$\mathcal{L}\{y(t)\} = \frac{1}{s}\mathcal{L}\{f(t)\}$$

In a similar manner if $y() = \int_0^t \int_0^t f(t)dtdt$ then

$$\mathcal{L}\{y(t)\} = \frac{1}{s^2}\mathcal{L}\{f(t)\}.$$

This is inverted by convolution theorem to give

$$y(t) = \int_0^t (t-\tau)f(\tau)d\tau.$$

If $$y(t) = \underbrace{\int_0^t \int_0^t \cdots \int_0^t f(t)dtdt \cdots dt}_{n\text{-fold integrals}}$$

Then $\mathcal{L}\{y(t)\} = \frac{1}{s^n}\mathcal{L}\{f(t)\}$. Using the convolution theorem, we get the Laplace inverse as

$$y(t) = \int_0^t \frac{(t-\tau)^{n-1}}{(n-1)!} f(\tau)d\tau.$$

Thus the $n$-fold integrals can be expressed as a single integral as,

$$\underbrace{\int_0^t \int_0^t \cdots \int_0^t f(t)dtdt \cdots dt}_{n\text{-fold integrals}} = \int_0^t \frac{(t-\tau)^{n-1}}{(n-1)!} f(\tau)d\tau.$$

**Method of Successive Approximation to Solve Volterra Integral Equations of Second Kind**

Volterra integral equation of the second kind is of the form,

$$u(x) = f(x) + \lambda \int_0^x K(x,t)u(t)dt$$

where $K(x, t)$ is the kernel of the integral equation, $f(x)$ a continuous function of $x$ and $\lambda$ a parameter. Here, $f(x)$ and $K(x, t)$ are the given functions but $u(x)$ is an unknown function that needs to be determined. The limits of integral for the Volterra integral equations are functions of $x$.

In this method of approximation, we replace the unknown function $u(x)$ under the integral sign of the Volterra equation by any selective real valued continuous function $u_0(x)$, called the zeroth approximation. This substitution will give the first approximation $u_1(x)$ by

$$u_1(x) = f(x) + \lambda \int_0^x K(x,t)u_0(t)dt$$

It is obvious that $u_1(x)$ is continuous if $f(x)$, $K(x, t)$ and $u_0(x)$ are continuous. The second approximation $u_2(x)$ can be obtained similarly by replacing $u_0(x)$ in the above equation by $u_1(x)$ obtained above. And we find,

$$u_2(x) = f(x) + \lambda \int_0^x K(x,t)u_1(t)dt$$

Proceeding in a similar way, we obtain an infinite sequence of functions

$$u_0(x), u_1(x), u_2(x), \ldots, u_n(x), \ldots$$

that satisfies the recurrence relation

$$u_n(x) = f(x) + \lambda \int_0^x K(x,t)u_{n-1}(t)dt$$

for $n = 1, 2, 3, \ldots$ and $u_0(x)$ is equivalent to any selected real valued function. The most commonly selected function for $u_0(x)$ are 0, 1 and $x$. Thus, at the limit, the solution $u(x)$ of the Volterra equation is obtained as,

$$u(x) = \lim_{n\to\infty} u_n(x),$$

so that the resulting solution $u(x)$ is independent of the choice of the zeroth approximation $u_0(x)$. This process of approximation is extremely simple. However, if we follow the Picard's successive approximation method, we need to set $u_0(x) = f(x)$, and determine $u_1(x)$ and other successive approximation as follows:

**NOTES**

$$u_1(x) = f(x) + \lambda \int_0^x K(x,t)f(t)dt$$

$$u_2(x) = f(x) + \lambda \int_0^x K(x,t)u_1(t)dt$$

$$\cdots\cdots\cdots$$

$$u_{n-1}(x) = f(x) + \lambda \int_0^x K(x,t)u_{n-2}(t)dt$$

$$u_n(x) = f(x) + \lambda \int_0^x K(x,t)u_{n-1}(t)dt$$

The last equation is the recurrence relation. Consider

$$u_2(x) - u_1(x) = \lambda \int_0^x K(x,t)[f(t) + \lambda \int_0^t K(t,\tau)f(\tau)d\tau]dt$$

$$-\lambda \int_0^x K(x,t)f(t)dt$$

$$= \lambda^2 \int_0^x K(x,t) \int_0^t K(t,\tau)f(\tau)d\tau dt$$

$$= \lambda^2 \psi_2(x)$$

Where,

$$\psi_2(x) = \int_0^x K(x,t)dt \int_0^t K(t,\tau)f(\tau)d\tau$$

Thus, it can be easily observed that,

$$u_n(x) = \sum_{m=0}^{n} \lambda^m \psi_m(x)$$

if $\psi_0(x)=f(x)$, and further that

$$\psi_m(x) = \int_0^x K(x,t)\psi_{m-1}(t)dt,$$

where $m$=1, 2, 3, . . . and hence,

$\psi_1(x) = \int_0^x K(x,t)f(t)dt.$

NOTES

The repeated integrals in $\psi_2(x)$ may be considered as a double integral over the triangular region; thus interchanging the order of integration, we obtain

$$\psi_2(x) = \int_0^x f(\tau)d\tau \int_\tau^x K(x,t)K(t,\tau)dt$$
$$= \int_0^x K_2(x,\tau)f(\tau)d\tau$$

Where,

$K_2(x,\tau) = \int_\tau^x K(x,t)K(t,\tau)dt.$

Similarly,

$$\psi_m(x) = \int_0^x K_m(x,\tau)f(\tau)d\tau, \quad m = 1, 2, 3, \ldots$$

Where the iterative kernels, $K_1(x,t) \equiv K(x,t),\ K_2(x,t),\ K_3(x,t), \ldots$ are defined by the recurrence formula given by,

$$K_{m+1}(x,t) = \int_t^x K(x,\tau)K_m(\tau,t)d\tau, \quad m = 1, 2, 3, \ldots$$

Thus, the solution for $u_n(x)$ can be written as,

$$u_n(x) = f(x) + \sum_{m=1}^{n} \lambda^m \psi_m(x)$$

**Resolvent Kernel as a Series in L**

It is also possible that we should be led to the solution of Volterra equation by means of the sum if it exists, of the infinite series defined by $u_n(x)$. Thus, we have using $\psi_m(x)$

$$u_n(x) = f(x) + \sum_{m=1}^{n} \lambda^m \int_0^x K_m(x,\tau)f(\tau)d\tau$$
$$= f(x) + \int_0^x \left\{ \sum_{m=1}^{n} \lambda^m K_m(x,\tau) \right\} f(\tau)d\tau;$$

hence it is also possible that the solution of Volterra equation will be given by as $n \to \infty$

**NOTES**

$$\lim_{n\to\infty} u_n(x) = u(x)$$

$$= f(x) + \int_0^x \left\{ \sum_{m=1}^{n} \lambda^m K_m(x,\tau) \right\} f(\tau) d\tau$$

$$= f(x) + \lambda \int_0^x H(x,\tau;\lambda) f(\tau) d\tau$$

Where,

$$H(x,\tau;\lambda) = \sum_{m=1}^{n} \lambda^m K_m(x,\tau)$$

is known as the resolvent kernel.

**Laplace Transform Method for A Difference Kernel**

Volterra integral equation of convolution type such as

$$u(x) = f(x) + \lambda \int_0^x K(x-t) u(t) dt$$

where the kernel $K(x–t)$ is of convolution type, can very easily be solved using the Laplace transform method. To begin the solution process, we first define the Laplace transform of $u(x)$

$$\mathcal{L}\{u(x)\} = \int_0^\infty e^{-sx} u(x) dx.$$

Using the Laplace transform of the convolution integral, we have

$$\mathcal{L}\left\{ \int_0^x K(x-t) u(t) dt \right\} = \mathcal{L}\{K(x)\}\mathcal{L}\{u(x)\}$$

Thus, taking the Laplace transform of Volterra integral equations of convolution type, we obtain

$$\mathcal{L}\{u(x)\} = \mathcal{L}\{f(x)\} + \lambda \mathcal{L}\{K(x)\}\mathcal{L}\{u(x)\}$$

and the solution for $L\{u(x)\}$ is given by

$$\mathcal{L}\{u(x)\} = \frac{\mathcal{L}\{f(x)\}}{1 - \lambda \mathcal{L}\{K(x)\}},$$

By inverting this transform, we obtain

$$u(x) = \int_0^x \psi(x-t) f(t) dt \text{ where } \mathcal{L}^{-1}\left\{ \frac{1}{1-\lambda\mathcal{L}\{K(x)\}} \right\} = \psi(x)$$

**NOTES**

**Example 1:** Solve the following Volterra integral equation of the second kind of the convolution type using (a) the Laplace transform method and (b) successive approximation method

$$u(x) = f(x) + \lambda \int_0^x e^{x-t} u(t)dt$$

**Solution:** (a) Taking the Laplace transforms, we obtain

$$\mathcal{L}\{u(x)\} = \mathcal{L}\{f(x)\} + \lambda \mathcal{L}\{e^x\}\mathcal{L}\{u(x)\},$$

and solving for $L\{u(x)\}$ yields

$$\mathcal{L}\{u(x)\} = \left(1 + \frac{\lambda}{s-1-\lambda}\right)\mathcal{L}\{f(x)\}.$$

The Laplace inverse of the above can be written immediately as

$$u(x) = \int_0^x \{\delta(x-t) + \lambda e^{(1+\lambda)(x-t)}\} f(t)dt$$

$$= f(x) + \lambda \int_0^x e^{(1+\lambda)(x-t)} f(t)dt$$

where $\delta(x)$ is the Dirac delta.

(b) Solution by successive approximation

Let us assume that the zeroth approximation is,

$u_0(x) = 0$

Then the first approximation can be obtained as

$u_1(x) = f(x)$

Hence, the second approximation is given by

$$u_2(x) = f(x) + \lambda \int_0^x e^{x-t} f(t)dt$$

Proceeding in similar manner, the third approximation can be obtained as

$$u_3(x) = f(x) + \lambda \int_0^x e^{x-t} u_2(t)dt$$

$$= f(x) + \lambda \int_0^x e^{x-t}\left\{f(t) + \lambda \int_0^t e^{t-\tau} f(\tau)d\tau\right\} dt$$

$$= f(x) + \lambda \int_0^x e^{x-t} f(t)dt + \lambda^2 \int_0^x \int_0^t e^{x-\tau} f(\tau)d\tau dt$$

$$= f(x) + \lambda \int_0^x e^{x-t} f(t)dt + \lambda^2 \int_0^x (x-t)e^{x-t} f(t)dt$$

**NOTES**

In the double integration the order of integration is changed to obtain the final result. In a similar manner, the fourth approximation $u_4(x)$ can be at once written as

$$u_4(x) = f(x) + \lambda \int_0^x e^{x-t} f(t)dt + \lambda^2 \int_0^x (x-t)e^{x-t} f(t)dt$$
$$+ \lambda^3 \int_0^x \frac{(x-t)^2}{2!} e^{x-t} f(t)dt.$$

Now, as $n \to \infty$

$$u(x) = \lim_{n\to\infty} u_n(x)$$
$$= f(x) + \lambda \left\{ \int_0^x e^{x-t} \left( 1 + \lambda(x-t) + \frac{1}{2!}\lambda^2(x-t)^2 + \cdots \right) f(t)dt \right\}$$
$$= f(x) + \lambda \int_0^x e^{(x-t)}.e^{\lambda(x-t)} f(t)dt$$
$$= f(x) + \lambda \int_0^x e^{(1+\lambda)(x-t)} f(t)dt$$

Here, the resolvent kernel is $H(x, t; \lambda) = e^{(1+\lambda)(x-t)}$.

**Method of Successive Substitution to Solve Volterra Integral Equations of Second Kind**

In this method, we substitute successively for $u(x)$ its value as given by Volterra integral equation of the second kind. We find that

$$u(x) = f(x) + \lambda \int_0^x K(x,t) \left\{ f(t)\lambda \int_0^t K(t,t_1)u(t_1)dt_1 \right\} dt$$
$$= f(x) + \lambda \int_0^x K(x,t)f(t)dt + \lambda^2 \int_0^x K(x,t) \int_0^t K(t,t_1)u(t_1)dt_1 dt$$
$$= f(x) + \lambda \int_0^x K(x,t)f(t)dt + \lambda^2 \int_0^x K(x,t) \int_0^t K(t,t_1)f(t_1)dt_1 dt$$
$$+ \cdots$$
$$+ \lambda^n \int_0^x K(x,t) \int_0^t K(t,t_1) \cdots$$
$$\times \int_0^{t_{n-2}} K(t_{n-2},t_{n-1})f(t_{n-1})dt_{n-1} \cdots dt_1 dt + R_{n+1}(x)$$

Where,

$$R_{n+1} = \lambda^{n+1} \int_0^x K(x,t) \int_0^t K(t,t_1) \cdots \int_0^{t_{n-1}} K(t_{n-1},t_n)u(t_n)dt_n \cdots dt_1 dt$$

is the remainder after $n$ terms.

Now,

$\lim_{n\to\infty} R_{n+1} = 0$

**NOTES**

Accordingly, the general series for $u(x)$ can be written as

$$u(x) = f(x) + \lambda \int_0^x K(x,t)f(t)dt$$
$$+\lambda^2 \int_0^x \int_0^t K(x,t)K(t,t_1)f(t_1)dt_1 dt$$
$$+\lambda^3 \int_0^x \int_0^t \int_0^{t_1} K(x,t)K(t,t_1)K(t_1,t_2)f(t_2)dt_2 dt_1 dt$$
$$+\cdots$$

**Example 2:** Solve the integral $u(x) = 1 + \int_0^x u(t)dt.$

**Solution:** By the method of successive substitution, we get

$$u(x) = 1\int_0^x dt + \int_0^x \int_0^x dt^2 + \int_0^x \int_0^x \int_0^x dt^3 + \cdots$$
$$= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$
$$= e^x.$$

**Iterated Kernels and Neumann Series for Volterra Equations**

The integral,

$u(x) = f(x) + \lambda \int_a^x \Gamma(x,\xi;\lambda)f(\xi)d\xi$ where $\Gamma(x,\xi;\lambda)$ is the resolvent kernel is the solution of the Volterra integral equation of the second kind, given by

$$u(x) = f(x) + \lambda \int_a^x K(x,\xi)u(\xi)d\xi$$

When both $K(x,\xi)$ and $f(x)$ are continuous then the resolvent kernel can be constructed in terms of the Neumann series

$$\Gamma(x,\xi;\lambda) = \sum_{n=0}^{\infty} \lambda^n K_{n+1}(x,\xi)$$

Where $K_{n+1}(x,\xi)$ is the iterated kernel which is evaluated as,

$$K_{n+1}(x,\xi) = \int_{\xi}^{x} K(x,y)K_n(y,\xi)dy$$

and $K_1(x,y) \equiv K(x,y)$.

For showing this, assume the following infinite series form for the solution $u(x)$,

$$u(x) = u_0(x) + \lambda u_1(x) + \lambda^2 u_2(x) + \cdots$$

Substituting this in the Volterra integral equation of the second kind and assuming good convergence which allows the exchange of summation with the integration operation, we get

$$u_0(x) + \lambda u_1(x) + \lambda^2 u_2(x) + \cdots$$
$$= f(x) + \lambda \int_a^x K(x,\xi)u_0(\xi)d\xi + \lambda^2 \int_a^x K(x,\xi)u_1(\xi)d\xi + \cdots$$

Equating coefficients of $\lambda$ on both sides, we have

$$u_0(x) = f(x)$$
$$u_1(x) = \int_a^x K(x,\xi)u_0(\xi)d\xi$$
$$u_2(x) = \int_a^x K(x,\xi)u_1(\xi)d\xi$$
$$\vdots$$
$$u_n(x) = \int_a^x K(x,\xi)u_{n-1}(\xi)d\xi.$$

By successive substitution, we get

$$u_1(x) = \int_a^x K(x,\xi)f(\xi)d\xi$$

And

$$u_2(x) = \int_a^x K(x,\xi)\int_a^{\xi} K(\xi,t)f(t)dtd\xi$$

**NOTES**

$$u_2(x) = \int_a^x f(t)\left[\int_t^x K(x,\xi)K(\xi,t)d\xi\right]dt$$

$$u_2(x) = \int_a^x f(t)K_2(x,t)dt$$

$$= \int_a^x K_2(x,\xi)f(\xi)d\xi$$

Similarly,

$$K_2(x,t) = \int_t^x K(x,\xi)K(\xi,t)d\xi$$

$$= \int_t^x K(x,\xi)K_1(\xi,t)d\xi$$

So we can now write,

$K_{n+1}(x,t) = \int_t^x K(x,\xi)K_n(\xi,t)d\xi,$ as the general term of the iterated kernel and

$$u_{n+1}(x) = \int_a^x K_{n+1}(x,\xi)f(\xi)d\xi.$$

Therefore,

$$\begin{aligned} u(x) &= f(x) + \lambda\int_a^x K_1(x,\xi)f(\xi)d\xi + \lambda^2\int_a^x K_2(x,\xi)f(\xi)d\xi \\ &\quad + \cdots + \lambda^n\int_a^x K_n(x,\xi)f(\xi)d\xi + \cdots \\ &= f(x) + \lambda\int_a^x [K_1(x,\xi) + \lambda K_2(x,\xi) + \cdots \\ &\quad + \lambda^n K_n(x,\xi) + \cdots]f(\xi)d\xi \\ &= f(x) + \lambda\int_a^x \left[\sum_{n=0}^{\infty}\lambda^n K_{n+1}(x,\xi)\right] f(\xi)d\xi \\ &= f(x) + \lambda\int_a^x \Gamma(x,\xi;\lambda)f(\xi)d\xi \end{aligned}$$

### Solution of a Volterra Integral Equation of the First Kind

The first kind Volterra equation is usually written as,

NOTES

$$\int_0^x K(x,t)u(t)dt = f(x)$$

If the derivatives $\frac{df}{dx} = f'(x)$, $\frac{\partial K}{\partial x} = K_x(x,t)$, and $\frac{\partial K}{\partial t} = K_t(x,t)$ exist and are continuous, then the solution of this equation is found by reducing it to its second kind and then proceeding with the methods discussed above.

Differentiating the above Volterra equation and applying Leibnitz rule, we get

$$K(x,x)u(x) + \int_0^x k_x(x,t)u(t)dt = f'(x)$$

If $K(x,x) \neq 0$ then

$$K(x,x)u(x) + \int_0^x k_x(x,t)u(t)dt = f'(x)$$

$$u(x) + \int_0^x \frac{k_x(x,t)}{K(x,x)}u(t)dt = \frac{f'(x)}{K(x,x)}$$

The second way to obtain the second kind Volterra integral equation from the first kind is by using integration by parts, if we set

$$\int_0^x u(t)dt = \phi(x)$$

Or $\int_0^t u(\xi)d\xi = \phi(t)$

By integrating by parts, we have

$$\left[K(x,t)\int_0^t u(\xi)d\xi\right]_{t=0}^x - \int_0^x K_t(x,t)\left(\int_0^t u(\xi)d\xi\right)dt = f(x)$$

which reduces to $[K(x,t)\phi(t)]_{t=0}^x - \int_0^x K_t(x,t)\phi(t)dt = f(x)$

Giving

$$K(x,x)\phi(x) - K(x,0)\phi(0) - \int_0^x K_t(x,t)\phi(t)dt = f(x)$$

$\varphi(0) = 0$, and dividing out by $K(x, x)$ we have

$$\phi(x) = \left\{\frac{f(x)}{K(x,x)}\right\} + \int_0^x \left\{\frac{K_t(x,t)}{K(x,x)}\right\} \phi(t)dt$$
$$= F(x) + \int_0^x G(x,t)\phi(t)dt$$

where $F(x) = \frac{f(x)}{K(x,x)}$ and $G(x,t) = \frac{K_t(x,t)}{K(x,x)}$.

**NOTES**

In this method the function $f(x)$ is not required to be differentiable. But $u(x)$ must finally be calculated by differentiating the function $\varphi(x)$ given by the formula

$$\phi(x) = \left\{\frac{f(x)}{K(x,x)}\right\} + \int_0^x H(x,t;1) \left\{\frac{f(t)}{K(t,t)}\right\} dt$$

where $H(x, t : 1)$ is the resolvent kernel corresponding to $\frac{K_t(x,t)}{K(x,x)}$. To do this $f(x)$ must be differentiable.

**Boundary Value Problems Reduced to Fredholm Integral Equations**

A boundary value problem can be converted to an equivalent Fredholm integral equation. But this method is complicated and so is rarely used. This method is demonstrated with the help of following illustration:

Consider the differential equation

$y''(x) + P(x)y'(x) + Q(x)y(x) = f(x)$ with boundary conditions

$$x = a : \quad y(a) = \alpha$$
$$y = b : \quad y(b) = \beta$$

Where $\alpha$ and $\beta$ are given constants. Make the transformation,

$$y''(x) = u(x)$$

Integrating both sides from $a$ to $x$, we get

$$y'(x) = y'(a) + \int_a^x u(t)dt$$

Integrating with respect to $x$ from $a$ to $x$ and applying the given boundary condition at $x = a$, we get

NOTES

$$y(x) = y(a) + (x-a)y'(a) + \int_a^x \int_a^x u(t)\,dt\,dt$$

$$= \alpha + (x-a)y'(a) + \int_a^x \int_a^x u(t)\,dt\,dt$$

And using the boundary condition at $x = b$ gives,

$$y(b) = \beta = \alpha + (b-a)y'(a) + \int_a^b \int_a^b u(t)\,dt\,dt,$$

And the unknown constant $y'(a)$ is determined as

$$y'(a) = \frac{\beta - \alpha}{b-a} - \frac{1}{b-a}\int_a^b \int_a^b u(t)\,dt\,dt.$$

Hence the solution can be rewritten as,

$$y(x) = \alpha + (x-a)\left\{\frac{\beta - \alpha}{b-a} - \frac{1}{b-a}\int_a^b \int_a^b u(t)\,dt\,dt\right\}$$
$$+ \int_a^x \int_a^x u(t)\,dt\,dt$$

Therefore,

$$u(x) = f(x) - P(x)\left\{y'(a) + \int_a^x u(t)\,dt\right\}$$
$$- Q(x)\left\{\alpha + (x-a)y'(a) + \int_a^x \int_a^x u(t)\,dt\,dt\right\}$$

where $u(x) = y''(x)$ and so $y(x)$ can be determined. It is a complicated procedure to determine the solution of a Boundary Value Problem (BVP) by equivalent Fredholm equation.

If $a = 0$ and $b = 1$, i.e., $0 \le x \le 1$, then

**NOTES**

$$y(x) = \alpha + xy'(0) + \int_0^x \int_0^x u(t)\,dt\,dt$$

$$= \alpha + xy'(a) + \int_0^x (x - t)u(t)\,dt$$

And hence the unknown constant $y'(0)$ can be determined as

$$y'(0) = (\beta - c) - \int_0^1 (1 - t)u(\ )dt$$

$$= (\beta - \alpha) - \int_0^x (1 - t)u(\ )dt - \int_x^1 (1 - t)u(\ )dt$$

Thus,

$$u(x) = f(x) - P(x) \left\{ y'(0) + \int_0^x u(t)\,dt \right\}$$

$$- Q(x)\left\{ \alpha + xy'(0) + \int_0^x (x - t)u(t)\,dt \right\}$$

$$u(x) = f(x) - (\beta - c)(P(x) + xQ(x)) - \alpha Q(x) + \int_0^1 K(x,t)u(t)\,dt$$

Where the kernel $K(x,t)$ is given by,

$$K(x,t) = \begin{cases} (P(x) + tQ(x))(1 - x) & 0 \le t \le x \\ (P(x) + xQ(x))(1 - t) & x \le t \le 1 \end{cases}$$

It can be easily verified that $K(x,t) = K(t,x)$ confirming that the kernel is symmetric. The Fredholm integral equation is given by $u(x)$.

**NOTES**

**Example 3:** Consider the boundary value problem,

$$y''(x) = f(x, y(x)), \quad 0 \le x \le 1$$

$$y(0) = y_0, \quad y(1) = y_1$$

**Solution:** Integrating the equation with respect to $x$ from 0 to $x$ two times yields

$$y(x) = y(0) + xy'(0) + \int_0^x \int_0^x f(t, y(t))dtdt$$

$$= y_0 + xy'(0) + \int_0^x (x - t)f(t, y(t))dt$$

To determine the unknown constant $y'(0)$, we use the condition at $x = 1$, i.e., $y(1) = y_1$. Hence,

$$y(1) = y_1 = y_0 + y'(0) + \int_0^1 (1 - t)f(t, y(t))dt,$$

And

$$y'(0) = (y_1 - y_0) - \int_0^1 (1 - t)f(t, y(t))dt.$$

Therefore,

$$y(x) = y_0 + x(y_1 - y_0) - \int_0^1 K(x, t)f(t, y(t))dt, \quad 0 \le x \le 1$$

Where the kernel is given by,

$$K(x, t) = \begin{cases} t(1 - t) & 0 \le t \le x \\ x(1 - t) & x \le t \le 1. \end{cases}$$

If we specialize our problem with simple linear BVP $y''(x) = -\lambda y(x)$, $0 < x < 1$ with the boundary conditions $y(0) = y_0$, $y(1) = y_1$, then $y(x)$ reduces to the second kind Fredholm integral equation,

$$y(x) = F(x) + \lambda \int_0^1 K(x, t)y(t)dt, \quad 0 \le x \le 1$$

where $F(x) = y_0 + x(y_1 - y_0)$. It can be easily verified that $K(x, t) = K(t, x)$ confirming that the kernel is symmetric.

NOTES

**Method of Successive Approximation to Solve Fredholm Equations of Second Kind**

The successive approximation method, which was successfully applied to Volterra integral equations of the second kind, can also be applied to the basic Fredholm integral equations of the second kind:

$$u(x) = f(x) + \lambda \int_a^b K(x,t)u(t)dt$$

We set $u_0(x) = f(x)$. Note that the zeroth approximation can be any selected real valued function $u_0(x)$, $a \le x \le b$.

Accordingly, the first approximation $u_1(x)$ of the solution of $u(x)$ is defined by

$$u_1(x) = f(x) + \lambda \int_a^b K(x,t)u_0(t)dt$$

The second approximation $u_2(x)$ of the solution $u(x)$ can be obtained by replacing $u_0(x)$ by the previously obtained $u_1(x)$. Hence we find

$$u_2(x) = f(x) + \lambda \int_a^b K(x,t)u_1(t)dt.$$

This process can be continued in the same manner to obtain the $n$th approximation. In other words, the various approximations can be put in a recursive scheme given by

$$u_0(x) = \text{any selective real valued function}$$
$$u_n(x) = f(x) + \lambda \int_a^b K(x,t)u_{n-1}dt, \quad n \ge 1.$$

Even though we can select any real valued function for the zeroth approximation $u_0(x)$, the most commonly selected functions for $u_0(x)$ are $u_0(x) = 0$, 1 or $x$. With the selection of $u_0(x) = 0$, the first approximation $u_1(x) = f(x)$. The final solution $u(x)$ is obtained by

$$u(x) = \lim_{n\to\infty} u_n(x)$$

so that the resulting solution $u(x)$ is independent of the choice of $u_0(x)$. This is known as Picard's method. The Neumann series is obtained if we set $u_0(x) = f(x)$ such that

$$u_1(x) = f(x) + \lambda \int_a^b K(x,t)u_0(t)dt$$
$$= f(x) + \lambda \int_a^b K(x,t)f(t)dt$$
$$= f(x) + \lambda\psi_1(x)$$

NOTES

Where,

$$\psi_1(x) = \int_a^b K(x,t)f(t)dt$$

The second approximation $u_2(x)$ can be obtained as,

$$u_2(x) = f(x) + \lambda \int_a^b K(x,t)u_1(t)dt$$

$$= f(x)\lambda \int_a^b K(x,t)\{f(t) + \lambda\psi_1(t)\}\,dt$$

$$= f(x) + \lambda\psi_1(x) + \lambda^2\psi_2(x)$$

Where,

$$\psi_2(x) = \int_a^b K(x,t)\psi_1(t)dt$$

The final solution $u(x)$ known as Neumann series can be obtained as,

$$u(x) = f(x) + \lambda\psi_1(x) + \lambda^2\psi_2(x) + \cdots + \lambda^n\psi_n(x) + \cdots$$

$$= f(x) + \sum_{n=1}^{\infty} \lambda^n \psi_n(x),$$

Where,

$$\psi_n(x) = \int_a^b K(x,t)\psi_{n-1}dt \quad n \geq 1$$

**Example 4:** Solve the Fredholm integral equation

$$u(x) = 1 + \int_0^1 xu(t)dt$$

by using the successive approximation method.

**Solution:** Let us consider the zeroth approximation is $u_0(x) = 1$, and then the first approximation can be computed as

$$u_1(x) = 1 + \int_0^1 xu_0(t)dt$$

$$= 1 + \int_0^1 xdt$$

$$= 1 + x$$

NOTES

$$u_2(x) = 1 + \int_0^1 xu_1(t)dt$$
$$= 1 + \int_0^1 x(1+t)dt$$
$$= 1 + x\left(1 + \frac{1}{2}\right)$$

$$u_3(x) = 1 + x\int_0^1 \left(1 + \frac{3t}{2}\right)dt$$
$$= 1 + x\left(1 + \frac{1}{2} + \frac{1}{4}\right)$$

Thus,

$$u_n(x) = 1 + x\left\{1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{n-1}}\right\}$$

And

$$u(x) = \lim_{n\to\infty} u_n(x)$$
$$= 1 + \lim_{n\to\infty} x\sum_{k=0}^{n}\frac{1}{2^k}$$
$$= 1 + x\left(1 - \frac{1}{2}\right)^{-1}$$
$$= 1 + 2x \qquad \text{is the solution.}$$

### Method of Successive Substitutions to Solve Fredholm Equations of Second Kind

This method is almost analogous to the successive approximation method except that it concerns with the solution of the integral equation in a series form through evaluating single and multiple integrals.

$$u(x) = f(x) + \lambda\int_a^b K(x,t)u(t)dt$$

$K(x, t) \neq 0$, is real and continuous in the rectangle $R$, for which $a \leq x \leq b$ and $a \leq t \leq b$; $f(x) \neq 0$ is real and continuous in the interval $I$, for which $a \leq x \leq b$; and $\lambda$, a constant parameter.

NOTES

Substituting the value of $u(t)$ in this equation, we get

$$u(x) = f(x) + \lambda \int_a^b K(x,t)f(t)dt + \lambda^2 \int_a^b K(x,t) \int_a^b K(t,t_1)u(t_1)dt_1 dt$$

or

$$u(x) = f(x) + \lambda \int_a^b K(x,t)f(t)dt$$
$$+ \lambda^2 \int_a^b K(x,t) \int_a^b K(t,t_1)f(t_1)dt_1 dt$$
$$+ \lambda^3 \int_a^b K(x,t) \int_a^b K(t,t_1) \int_a^b K(t_1,t_2)u(t_2)dt_2 dt_1 dt$$

Hence,

$$u(x) = f(x) + \lambda \int_a^b K(x,t)f(t)dt$$
$$+ \lambda^2 \int_a^b K(x,t) \int_a^b K(t,t_1)f(t_1)dt_1 dt$$
$$+ \lambda^3 \int_a^b K(x,t) \int_a^b K(t,t_1) \int_a^b K(t_1,t_2)f(t_2)dt_2 dt_1 dt$$
$$+ \cdots$$

The unknown function $u(x)$ is replaced by the known function $f(x)$.

**Example 5:** Use the successive substitutions to solve the Fredholm integral equation

$$u(x) = \cos x + \frac{1}{2} \int_0^{\pi/2} \sin x u(t)dt.$$

**Solution:** Here, $\lambda = 12$, $f(x) = \cos x$, and $K(x, t) = \sin x$

$$u(x) = \cos x + \frac{1}{2} \int_0^{\pi/2} \sin x \cos t dt + \frac{1}{4} \int_0^{\pi/2} \sin x \int_0^{\pi/2} \sin t \cos t_1 dt_1 dt$$
$$+ \frac{1}{8} \int_0^{\pi/2} \sin x \int_0^{\pi/2} \sin t \int_0^{\pi/2} \sin t_1 \cos t_2 dt_2 dt_1 dt + \cdots$$
$$= \cos x + \frac{1}{2} \sin x + \frac{1}{4} \sin x + \frac{1}{8} \sin x + \cdots$$
$$= \cos x + \sin x$$

### Iterated Kernels and Neumann Series for Fredholm Equations

The Liouville-Neumann series is defined as,

NOTES

$$\phi(x) = \sum_{n=0}^{\infty} \lambda^n \phi_n(x)$$

It is a unique continuous solution of a Fredholm integral equation of the second kind, defined as

$$f(t) = \phi(t) - \lambda \int_a^b K(t,s)\phi(s)\,ds$$

If the $n$th iterated kernel is defined as

$$K_n(x,z) = \int\int\cdots\int K(x,y_1)\,K(y_1,y_2)\cdots$$

$$K(y_{n-1},z)\,dy_1 dy_2\cdots dy_{n-1}$$

Then,

$$\phi_n(x) = \int K_n(x,z)\,f(z)\,dz$$

And the resolvent kernel is given by

$$K(x,z;\lambda) = \sum_{n=0}^{\infty} \lambda^n K_{n+1}(x,z).$$

### Resolvent Kernel as a Sum Of Series

Let the Fredholm equation of the second kind be,

$$u(x) = h(x) + \int_a^b K(x,y)u(y)dy.$$

Where the range of the separable kernel

$$K(x,y) = \sum_{i=1}^{n} f_i(x)g_i(y).$$

which consists of arbitrary linear combinations of the functions $f_i$ is given by,

$$\int_a^b K(x,y)u(y)dy = \sum_{i=1}^{n} f_i(x)u_i.$$

NOTES

Therefore,

$$u(x) = h(x) + \sum_{j=1}^{n} f_j(x)u_j.$$

To find $u_j$ define,

$$h_i = \int_a^b g_i(x)h(x)dx$$

Consider the algebraic problem $\mathbf{Mu} = \mathbf{h}$, . If we replace $h$ by a kernel $cK$ instead of the separable $K$ then the equation becomes,

$$\mathbf{M} = \mathbf{I} - c\mathbf{K}$$

where, $I$ is the identity matrix. Let $D(c)$ be the determinant of the matrix $M$ then

$$D(c) = \det(\mathbf{I} - c\mathbf{K})$$

If determinant $D(c)$ is not zero then the matrix $M$ has an inverse,

$$\mathbf{M}^{-1} = \mathbf{C}^t(c)/D(c) = \mathbf{R}(c)$$

Then the solution of the algebraic equation becomes $\mathbf{u} = \mathbf{M}^{-1}\mathbf{h}$ or $\mathbf{u} = \mathbf{Rh}$.

Now by substituting these values of $u_i$ in the Fredholm integral equation and then expressing $h_i$ in terms of $h(x)$, we get

$$u(x) = h(x) + \int_a^b cR(x, y; c)h(y)dy,$$

Where, $R(x, y; c) = \sum_i \sum_j R_{ij}(c)f_i(x)g_j(y)$ is called the resolvent kernel.

### Fredholm Resolvent Kernel as a Ratio of Two Series

If $K(x, \xi)$ is a continuous kernel, not necessarily real then the resolvent kernel $\Gamma(x, \xi; \lambda)$ can be expressed as the ratio of two infinite series of powers of $\lambda$ such that both of these series converge for all values of $\lambda$.

Expressing the resolvent kernel as a ratio,

$$\Gamma(x, \xi; \lambda) = \frac{D(x, \xi; \lambda)}{\Delta(\lambda)}$$

Where,

$$D(x, \xi; \lambda) = K(x, \xi) + \lambda D_1(x, \xi) + \lambda^2 D_2(x, \xi) + \cdots$$

And $\quad \Delta(\lambda) = 1 + \lambda C_1 + \lambda^2 C_2 + \cdots,$

Here the coefficients $C_i$ and the function $D_i(x, \xi)$ can be determined by,

$$C_1 = -\int_a K(x, x)\, dx, \quad D_1(x, \xi) = C_1 K(x, \xi) + \int_a^b K(x, \xi_1) K(\xi_1, \xi)\, d\xi_1;$$

$$2C_2 = -\int_a^b D_1(x, x)\, dx, \quad D_2(x, \xi) = C_2 K(x, \xi) + \int_a^b K(x, \xi_1) D_1(\xi_1, \xi)\, d\xi_1;$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$i C_i = -\int_a^b D_{i-1}(x, x)\, dx, \; D_i(x, \xi) = C_i K(x, \xi) + \int_a^b K(x, \xi_1) D_{i-1}(\xi_1, \xi)\, d\xi_1.$$

The solution of the equation given by,

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi$$

now becomes

$$y(x) = F(x) + \lambda \frac{\int_a^b D(x, \xi; \lambda) F(\xi)\, d\xi}{\Delta(\lambda)}.$$

This method is preferable only when the kernel is separable.

**Fredholm's Equations with Separable Kernels**

This section deals with the study of the homogeneous Fredholm integral equation with separable kernel given by,

$$u(x) = \lambda \int_0^b K(x, t) u(t) dt$$

This equation is obtained from the second kind Fredholm equation

$$u(x) = f(x) + \lambda \int_a^b K(x, t) u(t) dt,$$

Setting $f(x) = 0$, it is easily seen that $u(x) = 0$ is a solution which is known as the trivial solution. The homogeneous Fredholm integral equation with separable kernel may have nontrivial solutions. We shall use the direct computational method to obtain the solution in this case. Without loss of generality, we assume that

NOTES

$$K(x,t) = g(x)h(t)$$

So that,

$$u(x) = \lambda g(x)\int_a^b h(t)u(t)dt$$

For

$$\alpha = \int_a^b h(t)u(t)dt$$

$$u(x) = \lambda\alpha g(x)$$

We note that $\alpha = 0$ gives the trivial solution $u(x) = 0$. However, to determine the nontrivial solution, we need to determine the value of the parameter $\lambda$ by considering $\alpha \neq 0$. Therefore,

$$\alpha = \lambda\alpha\int_a^b h(t)g(t)dt$$

Or

$$1 = \lambda\int_a^b h(t)g(t)dt$$

which gives a numerical value for $\lambda \neq 0$ by evaluating the definite integral.

**Check Your Progress**

1. What is an integral equation?
2. Write the standard form of the Volterra equation.
3. What is first step in the method of successive approximation to solve Volterra equations?
4. Write the Volterra integral equation of convolution type.
5. What are the two methods to reduce a Volterra integral equation of the first kind to a second kind?
6. Express the resolvent kernel as a ratio of two series.

## 1.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. An integral equation is an equation in which an unknown function appears under an integral sign.

NOTES

2. The most standard form of Volterra linear integral equations is of the form

$$\phi(x)u(x) = f(x) + \lambda \int_a^x K(x,t)u(t)dt$$

3. In successive method of approximation, we first replace the unknown function $u(x)$ under the integral sign of the Volterra equation by any selective real valued continuous function $u_0(x)$.
4. Volterra integral equation of convolution type is

$$u(x) = f(x) + \lambda \int_0^x K(x-t)u(t)dt$$

where the kernel $K(x-t)$ is of convolution type.
5. The two methods to reduce a Volterra integral equation of the first kind to a second kind are differentiating the Volterra equation and applying Leibnitz rule, and by using integration by parts.
6. Expressing the resolvent kernel as a ratio,

$$\Gamma(x, \xi; \lambda) = \frac{D(x, \xi; \lambda)}{\Delta(\lambda)}$$

where, $D(x, \xi; \lambda) = K(x, \xi) + \lambda D_1(x, \xi) + \lambda^2 D_2(x, \xi) + \cdots$ and

$$\Delta(\lambda) = 1 + \lambda C_1 + \lambda^2 C_2 + \cdots,$$

## 1.4 SUMMARY

- An integral equation in $u(x)$ is given by,

$$u(x) = f(x) + \lambda \int_{\alpha(x)}^{\beta(x)} K(x,t)u(t)dt$$

where $K(x, t)$ is called the kernel of the integral equation, and $\alpha(x)$ and $\beta(x)$ are the limits of integration.
- The most frequently used integral equations fall under two major classes, namely Volterra and Fredholm integral equations.
- In Volterra equation one of the limits of integration is variable while in Fredholm equation both the limits are constant.

NOTES

## 1.5 KEY WORDS

- **Integral equation:** It is an equation in which an unknown function appears under an integral sign.

## 1.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. Write the two kinds of Volterra integral equations.
2. What is the basic difference between Volterra and Fredholm equations?
3. List the methods used to solve Fredholm and Volterra integral equations of the second kind.
4. How can you find the solution of the Volterra integral equation of the first kind?
5. Define iterated kernel for Fredholm and Volterra integral equations.

### Long-Answer Questions

1. Reduce the following initial value problem to an equivalent Volterra equation:

$$\frac{d^4y}{dx^4}+\frac{d^2y}{dx^2}=2e^x, \quad y(0)=2, y'(0)=2, y''(0)=1, y'''(0)=1.$$

2. Solve the following Volterra integral equations using methods of successive approximation with five approximations with $u_0(x)=0$:

$$(a) \quad u(x)=2x+2x^2-x^3+\int_0^x u(t)dt$$

$$(b) \quad u(x)=-1-\int_0^x u(t)dt$$

$$(c) \quad u(x)=1-x-\int_0^x (x-t)u(t)dt$$

$$(d) \quad u(x)=x\cos x+\int_0^x tu(t)dt.$$

3. Find the solution of the following Volterra integral equations of the first kind:

(a) $xe^{-x}=\int_0^x e^{t-x}u(t)dt$

NOTES

(b) $\tan x - \ln(\cos x) = \int_0^x (1 + x - t)u(t)dt, \quad x < \pi/2.$

(c) $5x^2 + x^3 = \int_0^x (5 + 3x - 3t)u(t)dt$

(d) $2\cosh x - \sinh x - (2 - x) = \int_0^x (2 - x + t)u(t)dt$

4. Reduce the following initial value problem into an equivalent Fredholm equation:

   $y'' + 2xy = 1, 0 < x < 1, \quad y(0) = 0, y(1) = 0$

5. Solve the following linear Fredholm integral equations:

   (a) $u(x) = \sec^2 x + \lambda \int_0^1 u(t)dt.$

   (b) $u(x) = \frac{5x}{6} + \frac{1}{2}\int_0^1 xtu(t)dt.$

   (c) $u(x) = \cos x + \lambda \int_0^{\pi} xtu(t)dt.$

   (d) $u(x) = \sec^2 x \tan x - \lambda \int_0^1 u(t)dt.$

   (e) $u(x) = e^x + \lambda \int_0^1 2e^x e^t u(t)dt.$

## 1.7 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

NOTES

# UNIT 2 METHODS FOR FINDING COMPLEX ROOTS AND POLYNOMIAL EQUATIONS

**Structure**



## 2.0 INTRODUCTION

In mathematics and computing, a root-finding algorithm is an algorithm for finding zeroes, also called roots, of continuous functions. A zero of a function $f$, from the real numbers to real numbers or from the complex numbers to the complex numbers, is a number $x$ such that $f(x) = 0$. As, generally, the zeroes of a function cannot be computed exactly nor expressed in closed form, root-finding algorithms provide approximations to zeroes, expressed either as floating point numbers or as small isolating intervals, or disks for complex roots (an interval or disk output being equivalent to an approximate output together with an error bound).

In this unit, you will study about the methods for finding complex roots and polynomial equations.

## 2.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the various methods for finding complex roots
- Analyse the polynomial equations

## 2.2 METHODS FOR FINDING COMPLEX ROOTS

In this section, we consider numerical methods for computing the roots of an equation of the form,

$$f(x) = 0 \qquad (2.1)$$

NOTES

Where $f(x)$ is a reasonably well-behaved function of a real variable $x$. The function may be in algebraic form or polynomial form given by,

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0 \quad (2.2)$$

It may also be an expression containing transcendental functions such as cos $x$, sin $x$, $e^x$, etc. First, we would discuss methods to find the isolated real roots of a single equation. Later, we would discuss methods to find the isolated roots of a system of equations, particularly of two real variables $x$ and $y$, given by,

$$f(x, y) = 0\ ,\ g(x, y) = 0 \quad (2.3)$$

A root of an equation is usually computed in two stages. First, we find the location of a root in the form of a crude approximation of the root. Next we use an iterative technique for computing a better value of the root to a desired accuracy in successive approximations/computations. This is done by using an iterative function.

### Methods for Finding Location of Real Roots

The location or crude approximation of a real root is determined by the use of any one of the following methods, (i) graphical and (ii) tabulation.

**Graphical Method:** In the graphical method, we draw the graph of the function $y = f(x)$, for a certain range of values of $x$. The abscissae of the points where the graph intersects the $x$-axis are crude approximations for the roots of the Equation (2.1). For example, consider the equation,

$$f(x) = x^2 + 2x - 1 = 0$$

From the graph of the function $y = f(x)$, shown in Figure 2.1 we find that it cuts the $x$-axis between 0 and 1. We may take any point in [0, 1] as the crude approximation for one root. Thus, we may take 0.5 as the location of a root. The other root lies between – 2 and – 3. We can take – 2.5 as the crude approximation of the other root.

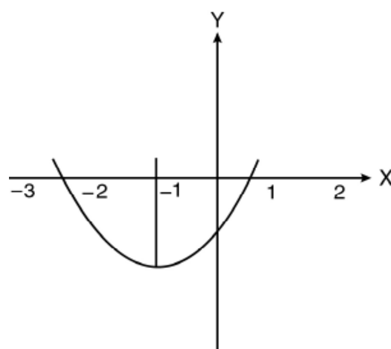



***Fig. 2.1*** *Graph of* $y = x^2 + 2x - 1$

In some cases, where it is complicated to draw the graph of $y = f(x)$, we may rewrite the equation $f(x) = 0$, as $f_1(x) = f_2(x)$, where the graphs of $y = f_1(x)$ and $y = f_2(x)$ are standard curves. Then we find the $x$-coordinate(s) of the point(s) of

NOTES

intersection of the curves $y = f_1(x)$, and $y = f_2(x)$, which is the crude approximations of the root (s).

For example, consider the equation,

$$x^3 - 15.2x - 13.2 = 0$$

This can be rewritten as,

$$x^3 = 15.2x + 13.2$$

Where it is easy to draw the graphs of $y = x^3$ and $y = 15.2\,x + 13.2$. Then, the abscissa of the point(s) of intersection can be taken as the crude approximation(s) of the root(s).

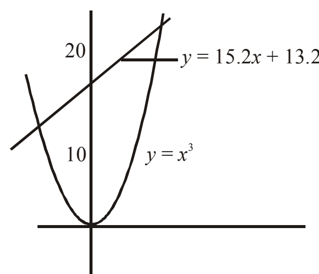



***Fig. 2.2*** *Graph of $y = x^3$ and $y = 15.2x + 13.2$*

**Example 1:** Find the location of the root of the equation $x \log_{10} x = 1$.

**Solution:** The equation can be rewritten as $\log_{10} x = \frac{1}{x}$.

Now the curves $y = \log_{10} x$, and $y = \frac{1}{x}$, can be easily drawn and are shown in Figure 2.3.

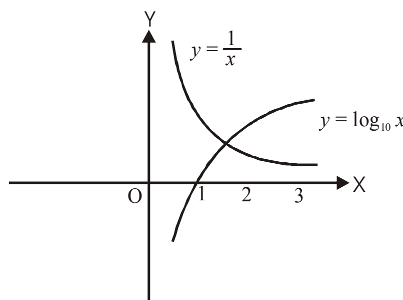



***Fig. 2.3*** *Graph of $y = \frac{1}{x}$ and $y = \log_{10} x$*

The point of intersection of the curves has its $x$-coordinates value 2.5 approximately. Thus, the location of the root is 2.5.

**Tabulation Method:** In the tabulation method, a table of values of $f(x)$ is made for values of $x$ in a particular range. Then, we look for the change in sign in the values of $f(x)$ for two consecutive values of $x$. We conclude that a real root lies

between these values of $x$. This is true if we make use of the following theorem on continuous functions.

**Theorem 1:** If $f(x)$ is continuous in an interval $(a, b)$ and $f(a)$ and $f(b)$ are of opposite signs, then there exists at least one real root of $f(x) = 0$, between $a$ and $b$.

Consider for example, the equation $f(x) = x^3 - 8x + 5 = 0$

Constructing the following table of $x$ and $f(x)$

| $x$ | $-4$ | $-3$ | $-2$ | $-1$ | $0$ | $1$ | $2$ | $3$ |
|---|---|---|---|---|---|---|---|---|
| $f(x)$ | $-27$ | $2$ | $13$ | $12$ | $5$ | $-2$ | $-3$ | $8$ |

We observe that there is a change in sign of $f(x)$ in each of the sub-intervals (–3, –4), (0, 1) and (2, 3). Thus we can take the crude approximation for the three real roots as – 3.2, 0.2 and 2.2.

### Methods for Finding the Roots—Bisection and Simple Iteration Methods

**Bisection Method:** The bisection method involves successive reduction of the interval in which an isolated root of an equation lies. This method is based upon an important theorem on continuous functions as stated below.

**Theorem 2:** If a function $f(x)$ is continuous in the closed interval $[a, b]$ and $f(a)$ and $f(b)$ are of opposite signs, i.e., $f(a)f(b) < 0$, then there exists at least one real root of $f(x) = 0$ between $a$ and $b$.

The bisection method starts with two guess values $x_0$ and $x_1$. Then this interval $[x_0, x_1]$ is bisected by a point $x_2 = \frac{1}{2}(x_0 + x_1)$, where $f(x_0) \cdot f(x_1) < 0$. We compute $f(x_2)$. If $f(x_2) = 0$, then $x_2$ is a root. Otherwise, we check whether $f(x_0) \cdot f(x_2) < 0$ or $f(x_1) \cdot f(x_2) < 0$. If $f(x_2)/f(x_0) < 0$, then the root lies in the interval $(x_2, x_0)$. Otherwise, if $f(x_0) \cdot f(x_1) < 0$, then the root lies in the interval $(x_2, x_1)$.

The sub-interval in which the root lies is again bisected and the above process is repeated until the length of the sub-interval is less than the desired accuracy.

The bisection method is also termed as a bracketing method, since the method successively reduces the gap between the two ends of an interval surrounding the real root, i.e., brackets the real root.

The algorithm given below clearly shows the steps to be followed in finding a real root of an equation, by bisection method to the desired accuracy.

**Algorithm:** Finding root using bisection method.

**Step 0:** Define the equation, $f(x) = 0$

**Step 1:** Read epsilon, the desired accuracy

**Setp 2:** Read two initial values $x_0$ and $x_1$ which bracket the desired root

**Step 3:** Compute $y_0 = f(x_0)$

**Step 4:** Compute $y_1 = f(x_1)$

NOTES

**Step 5:** Check if $y_0 y_1 < 0$, then go to Step 6
else go to Step 2

**Step 6:** Compute $x_2 = (x_0 + x_1)/2$

**Step 7:** Compute $y_2 = f(x_2)$

**Step 8:** Check if $y_0 y_2 > 0$, then set $x_0 = x_2$
else set $x_1 = x_2$

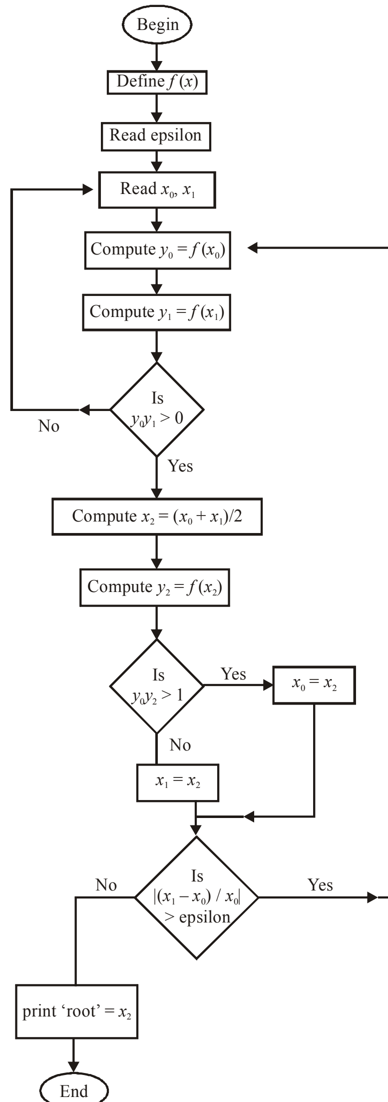
**Step 9:** Check if $|(x_1 - x_0)/x_1| >$ epsilon, then go to Step 3

**Step 10:** Write $x_2, y_2$

**Step 11:** End

Next, we give the flowchart representation of the above algorithm to get a better understanding of the method. The flowchart also helps in easy implementation of the method in a computer program.

**Flow Chart for Bisection Algorithm**



**Example 2:** Find the location of the smallest positive root of the equation $x^3 - 9x + 1 = 0$ and compute it by bisection method, correct to two decimal places.

**Solution:** To find the location of the smallest positive root we tabulate the function $f(x) = x^3 - 9x + 1$ below:

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $f(x)$ | 1 | −2 | −9 | 1 |

We observe that the smallest positive root lies in the interval [0, 1]. The computed values for the successive steps of the bisection method are given in the Table.

| $n$ | $x_0$ | $x_1$ | $x_2$ | $f(x_2)$ |
|---|---|---|---|---|
| 1 | 0 | 1 | 0.5 | − 3.37 |
| 2 | 0 | 0.5 | 0.25 | − 1.23 |
| 3 | 0 | 0.25 | 0.125 | − 0.123 |
| 4 | 0 | 0.125 | 0.0625 | 0.437 |
| 5 | 0.0625 | 0.125 | 0.09375 | 0.155 |
| 6 | 0.09375 | 0.125 | 0.109375 | 0.016933 |
| 7 | 0.109375 | 0.125 | 0.11718 | − 0.053 |

From the above results, we conclude that the smallest root correct to two decimal places is 0.11.

**Simple Iteration Method:** A root of an equation $f(x) = 0$, is determined using the method of simple iteration by successively computing better and better approximation of the root, by first rewriting the equation in the form,

$$x = g(x) \tag{2.4}$$

Then, we form the sequence $\{x_n\}$ starting from the guess value $x_0$ of the root and computing successively,

$$x_1 = g(x_0), \quad x_2 = g(x_1), .., \quad x_n = g(x_{n-1})$$

In general, the above sequence may converge to the root $\xi$ as $n \to \infty$, or it may diverge. If the sequence diverges, we shall discard it and consider another form $x = h(x)$, by rewriting $f(x) = 0$. It is always possible to get a convergent sequence since there are different ways of rewriting $f(x) = 0$, in the form $x = g(x)$. However, instead of starting computation of the sequence, we shall first test whether the form of $g(x)$ can give a convergent sequence or not. We give below a theorem which can be used to test for convergence.

**Theorem 3:** If the function $g(x)$ is continuous in the interval $[a, b]$ which contains a root $\xi$ of the equation $f(x) = 0$, and is rewritten as $x = g(x)$, and $|g'(x)| \leq l \leq 1$ in this interval, then for any choice of $x_0 \in [a, b]$, the sequence $\{x_n\}$ determined by the iterations,

NOTES

$$x_{k+1} = g(x_k), \text{ for } k = 0, 1, 2, ... \quad (2.5)$$

This converges to the root of $f(x) = 0$.

**Proof:** Since $x = \xi$, is a root of the equation $x = g(x)$, we have

$$\xi = g(\xi) \quad (2.6)$$

The first iteration gives $x_1 = g(x_0)$ (2.7)

Subtracting Equation (2.7) from Equation (2.6), we get

$$\xi - x_1 = g(\xi) - g(x_0)$$

Applying mean value theorem, we can write

$$\xi - x_1 = (\xi - x_0)g'(s_0), \quad x_0 < s_0 < \xi \quad (2.8)$$

Similarly, we can derive

$$\xi - x_2 = (\xi - x_1)g'(s_1), \quad x_1 < s_1 < \xi \quad (2.9)$$

....

$$\xi - x_{n+1} = (\xi - x_n)g'(s_n), \quad x_n < s_n < \xi \quad (2.10)$$

From all these Equations (2.8), (2.9), and (2.10), we get

$$\xi - x_{n+1} = (\xi - x_0)g'(s_0)g'(s_1)..., g'(s_n) \quad (2.11)$$

Since $|g'(x_i)| < l$ for each $x_i$, the above Equation (2.11) becomes,

$$|\xi - x_{n+1}| < l^{n+1} |\xi - x_0| \quad (2.12)$$

Evidently, since $l<1$, $l^{n+1} \to 0$, as $n \to \infty$, the right hand side tends to zero and thus it follows that the sequence $\{x_n\}$ converges to the root $\xi$ if $|\varphi'(\xi)| < 1$. This completes the proof.

**Order of Convergence:** The order of convergence of an iterative process is determined in terms of the errors $e_n$ and $e_{n+1}$ in successive iterations. An iterative process is said to have $k$th order convergence if $\lim_{n\to\infty} \frac{e_{n+1}}{e_n^k} < M$, where $M$ is a finite number.

Roughly speaking, the error in any iteration is proportional to the $k$th power of the error in the previous iteration.

Evidently, the simple iteration discussed in this section has its order of convergence 1.

The above iteration is also termed as fixed point iteration since it determines the root as the fixed point of the mapping defined by $x = g(x)$.

**Algorithm:** Computation of a root of $f(x) = 0$ by linear iteration.

**Step 0:** Define $g(x)$, where $f(x) = 0$ is rewritten as $x = g(x)$

**NOTES**

**Step 1:** Input $x_0$, epsilon, *maxit*, where $x_0$ is the initial guess of root, epsilon is accuracy desired, *maxit* is the maximum number of iterations allowed.

**Step 2:** Set $i = 0$

**Step 3:** Set $x_1 = g(x_0)$

**Step 4:** Set $i = i + 1$

**Step 5:** Check, if $|(x_1 - x_0)/x_1| <$ epsilon, then print 'root is', $x_1$ else go to Step 6

**Step 6:** Check, if $i < n$, then set $x_0 = x_1$ and go to Step 3

**Step 7:** Write 'No convergence after', $n$, 'iterations'

**Step 8:** End

**Example 3:** In order to compute a real root of the equation $x^3 - x - 1 = 0$, near $x = 1$, by iteration, determine which of the following iterative functions can be used to give a convergent sequence.

(*i*) $x = x^3 - 1$ (*ii*) $x = \frac{x+1}{x^2}$ (*iii*) $x = \sqrt{\frac{x+1}{x}}$

**Solution:**

(*i*) For the form $x = x^3 - 1$, $g(x) = x^3 - 1$, and $g'(x) = 3x^2$. Hence, $|g'(x)| > 1$, for $x$ near 1. So, this form would not give a convergent sequence of iterations.

(*ii*) For the form $x = \frac{x+1}{x^2}$, $g(x) = \frac{x+1}{x^2}$. Thus, $g'(x) = -\frac{1}{x^2} - \frac{2}{x^3}$ and $|g'(1)| = 3 > 1$. Hence, this form also would not give a convergent sequence of iterations.

(*iii*) For the form, $g(x) = \sqrt{\frac{x+1}{x}}$, $g'(x) = \frac{1}{2}\left(\frac{x+1}{x}\right)^{-\frac{1}{2}} \cdot \left(-\frac{1}{x^2}\right)$.

$\therefore$ $|g'(1)| = \frac{1}{2\sqrt{2}} < 1$. Hence, the form $x = \sqrt{\frac{x+1}{x}}$ would give a convergent sequence of iterations.

**Example 4:** Compute the real root of the equation $x^3 + x^2 - 1 = 0$, correct to five significant digits, by iteration method.

**Solution:** The equation has a real root between 0 and 1 since $f(x) = x^3 + x^2 - 1$ has opposite signs at 0 and 1. For using iteration, we first rewrite the equation in the following different forms.

(*i*) $x = \frac{1}{x^2} - 1$ (*ii*) $x = \sqrt{\frac{1}{x} - 1}$ (*iii*) $x = \frac{1}{\sqrt{x+1}}$

NOTES

For the form (*i*), $g(x) = -1 + \frac{1}{x^2}$, $g'(x) = -\frac{2}{x^3}$ and for $x$ in (0, 1), $|g'(x)| > 1$.

So, this form is not suitable. For the form $g'(x) = \frac{1}{2} \cdot \frac{1}{\sqrt{\frac{1}{x} - 1}} \left( -\frac{1}{x^2} - 1 \right)$ (*ii*) and $|g'(x)| > 1$ for all $x$ in (0, 1). Finally, for the form (*iii*) $g'(x) = -\frac{1}{2} \cdot \frac{1}{(x+1)^{\frac{3}{2}}}$ and $|g'(x)| < 1$ for $x$ in (0, 1). Thus this form can be used to form a convergent sequence for finding the root.

We start the iteration $x = \frac{1}{\sqrt{1+x}}$ with $x_0 = 1$. The results of suecessive iterations are,

$x_1 = 0.70711$ $x_2 = 0.76537$ $x_3 = 0.75236$ $x_4 = 0.75541$
$x_5 = 0.75476$ $x_6 = 0.75490$ $x_7 = 0.75488$ $x_8 = 0.75488$

Thus, the root is 0.75488, correct to five significant digits.

**Example 5:** Compute the root of the equation $x^2 - x - 0.1 = 0$, which lies in (1,2), correct to five significant figures.

**Solution:** The equation is rewritten in the following form for computing the root by iteration,

$x = \sqrt{x + 0.1}$. Here, $g'(x) = \frac{1}{2\sqrt{x + 0.1}}$, and $|g'(x)| < 1$, for $x$ in (1, 2).

The results for successive iterations, taking $x_0 = 1$, are

$x_1 = 1.0488$ $x_2 = 1.0718$ $x_3 = 1.0825$
$x_4 = 1.0874$ $x_5 = 1.0897$.

Thus, the root is 1.09, correct to three significant figures.

**Example 6:** Solve the following equation for the root lying in (2, 4) by using the method of linear iteration $x^3 - 9x + 1 = 0$. Show that there are various ways of rewriting the equation in the form, $x = g(x)$ and choose the one which gives a convergent sequence for the root.

**Solution:** We can rewrite the equation in the following different forms:

(*i*) $x = \frac{1}{9}(x^3 + 1)$ (*ii*) $x = 9/x - \frac{1}{x^2}$ (*iii*) $x = \sqrt{9 - \frac{1}{x}}$

In case of (*i*), $g'(x) = \frac{1}{3}x^2$ and for $x$ in [2, 4], $|g'(x)| > 1$. Hence it will not give rise to a convergent sequence.

NOTES

In case of form (*ii*) $g'(x) = 2x - \frac{9}{x^2} + \frac{2}{x^3}$ and for $x$ in [2, 4], $|g'(x)| > 1$

In case of form (*iii*) $g'(x) = \left(9 - \frac{1}{x}\right)^{-\frac{1}{2}} \frac{1}{2x^2}$ and $|g'(x)| < 1$

Thus, the forms (*ii*) and (*iii*) would give convergent sequences for finding the root in [2, 3].

We start the iterations taking $x_0 = 2$ in the iteration scheme (*iii*). The result for successive iterations are,

$x_0 = 2.0$ $x_1 = 2.91548$ $x_4 = 2.94282.$

$x_2 = 2.94228$ $x_3 = 2.94281$

Thus, the root can be taken as 2.94281, correct to four decimal places.

**Newton-Raphson Method**

Newton-Raphson method is a widely used numerical method for finding a root of an equation $f(x) = 0$, to the desired accuracy. It is an iterative method which has a faster rate of convergence and is very useful when the expression for the derivative $f(x)$ is not complicated. To derive the formula for this method, we consider a Taylor's series expansion of $f(x_0 + h)$, $x_0$ being an initial guess of a root of $f(x) = 0$ and $h$ is a small correction to the root.

$$f(x_0 + h) = f(x_0) + h\ f'(x_0) + \frac{h^2}{2!} f''(x_0) + ...$$

Assuming $h$ to be small, we equate $f(x_0 + h)$ to 0 by neglecting square and higher powers of $h$.

$\therefore$ $$f(x_0) + h\ f'(x_0) = 0$$

Or, $$h = -\frac{f(x_0)}{f'(x_0)}$$

Thus, we can write an improved value of the root as,

$$x_1 = x_0 + h$$

i.e., $$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Successive approximations $x_2, x_3, ..., x_{n+1}$ can thus be written as,

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$$

NOTES

... ... ...

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (2.13)$$

If the sequence $\{x_n\}$ converges, we get the root.

**Algorithm:** Computation of a root of $f(x) = 0$ by Newton-Raphson method.

**Step 0:** Define $f(x)$, $f'(x)$

**Step 1:** Input $x_0$, epsilon, *maxit*

[$x_0$ is the initial guess of root, epsilon is the desired accuracy of the root and *maxit* is the maximum number of iterations allowed]

**Step 2:** Set $i = 0$

**Step 3:** Set $f_0 = f(x_0)$

**Step 4:** Compute $df_0 = f'(x_0)$

**Step 5:** Set $x_1 = x_0 - f_0/df_0$

**Step 6:** Set $i = i + 1$

**Step 7:** Check if $|(x_1 - x_0)\ |x_1| <$ epsilon, then print 'root is', $x_1$ and stop

else if $i < n$, then set $x_0 = x_1$ and go to Step 3

**Step 8:** Write 'Iterations do not converge'

**Step 9:** End

**Example 7:** Use Newton-Raphson method to compute the positive root of the equation $x^3 - 8x - 4 = 0$, correct to five significant digits.

**Solution:** Newton-Raphson iterative scheme is given by,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad \text{for} \quad n = 0, 1, 2, ...$$

For the given equation $f(x) = x^3 - 8x - 4$

First we find the location of the root by the method of tabulation. The table for $f(x)$ is,

| $x$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $f(x)$ | −4 | −13 | −12 | −1 | 28 |

Evidently, the positive root is near $x = 3$. We take $x_0 = 3$ in Newton-Raphson iterative scheme.

$$x_{n+1} = x_n - \frac{x_n^3 - 8x_n - 4}{3x_n^2 - 8}$$

We get, $x_1 = 3 - \dfrac{27 - 24 - 4}{27 - 8} = 3.0526$

Similarly, $x_2 = 3.05138$, and $x_3 = 3.05138$

Thus, the positive root is 3.0514, correct to five significant digits.

**Example 8:** Find a real root of the equation $x^3 + 7x^2 + 9 = 0$, correct to five significant digits.

**Solution:** First we find the location of the real root by tabulation. We observe that the real root is negative and since $f(-7) = 9 > 0$ and $f(-8) = -55 < 0$, a root lies between –7 and – 8.

For computing the root to the desired accuracy, we take $x_0 = -8$ and use Newton-Raphson iterative formula,

$$x_{n+1} = x_n - \frac{x_n^3 + 7x_n^2 + 9}{3x_n^2 + 14x_n}, \quad \text{for} \quad n = 0, 1, 2, ...$$

The successive iterations give,

$$x_1 = -7.3125$$
$$x_2 = -7.17966$$
$$x_3 = -7.17484$$
$$x_4 = -7.17483$$

Hence, the desired root is –7.1748, correct to five significant digits.

**Example 9:** For evaluating $\sqrt{a}$, deduce the iterative formula $x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$, by using Newton-Raphson scheme of iteration. Hence, evaluate $\sqrt{2}$ using this, correct to four significant digits.

**Solution:** We observe that $\sqrt{a}$ is the solution of the equation $x^2 - a = 0$.

Now, using $f(x) = x^2 - a$ in the Newton-Raphson iterative scheme

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n}$$

We have, $$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n}$$

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n}$$

i.e., $$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \quad \text{for} \quad n = 0, 1, 2,...$$

Now, for computing $\sqrt{2}$, we assume $x_0 = 1.4$. The successive iterations give,

$$x_1 = \frac{1}{2}\left(1.4 + \frac{2}{1.4}\right) = \frac{3.96}{2.8} = 1.414$$

$$x_2 = \frac{1}{2}\left(1.414 + \frac{2}{1.414}\right) = 1.41421$$

Hence, the value of $\sqrt{2}$ is 1.414 correct to four significant digits.

**Example 10:** Prove that $\sqrt[k]{a}$ can be computed by the iterative scheme,

$x_{n+1} = \frac{1}{k}\left[(k-1)x_n + \frac{a}{x_n^{k-1}}\right]$. Hence evaluate $\sqrt[3]{2}$, connect to five significant digits.

**Solution:** The value $\sqrt[k]{a}$ is the positive root of $x^k - a = 0$. Thus, the iterative scheme for evaluating $\sqrt[k]{a}$ is,

$$x_{n+1} = x_n - \frac{x_n^k - a}{kx_n^{k-1}}$$

or, $$x_{n+1} = \frac{1}{k}\left[(k-1)x_n + \frac{a}{x_n^{k-1}}\right], \quad \text{for} \quad n = 0, 1, 2, ...$$

Now, for evaluating $\sqrt[3]{2}$, we take $x_0 = 1.25$ and use the iterative formula,

$$x_{n+1} = \frac{1}{3}\left[2x_n + \frac{2}{x_n^2}\right]$$

We have, $$x_1 = \frac{1}{3}\left[1.25 \times 2 + \frac{2}{(1.25)^2}\right] = 1.26$$

$$x_2 = 1.259921, \quad x_3 = 1.259921$$

Hence, $\sqrt[3]{2} = 1.2599$, correct to five significant digits.

**Example 11:** Find by Newton-Raphson method, the real root of $3x - \cos x - 1 = 0$, correct to three significant figures.

**Solution:** The location of the real root of $f(x) = 3x - \cos x - 1 = 0$, is [0, 1] since $f(0) = -2$ and $f(1) > 0$.

We choose $x_0 = 0$, and use Newton-Raphson scheme of iteration.

$$x_{n+1} = x_n - \frac{3x_n - \cos\ x_n - 1}{3 + \sin\ x_n}, \quad n = 0, 1, 2, ...$$

The results for successive iterations are,

$$x_1 = 0.667,\ x_2 = 0.6075,\ x_3 = 0.6071$$

Thus, the root is 0.607 correct to three significant figures.

NOTES

**Example 12:** Find a real root of the equation $x^x + 2x - 6 = 0$, correct to four significant digits.

**Solution:** Taking $f(x) = x^x + 2x - 6$, we have $f(1) = -3 < 0$ and $f(2) = 2 > 0$. Thus, a root lies in [1, 2]. Choosing $x_0 = 2$, we use Newton-Raphson iterative scheme given by,

$$x_{n+1} = x_n - \frac{x_n^{x_n} + 2x_n - 6}{x_n^{x_n}(\log_e x_n + 1) + 2}, \quad \text{for} \quad n = 0, 1, 2, \ldots$$

The computed results for successive iterations are,

$$x_1 = 2 - \frac{4 + 4 - 6}{4 \times (\log_e 2x^2 + 1) + 2} = 1.72238$$

$$x_2 = 1.72321$$

$$x_3 = 1.72308$$

Hence, the root is 1.723 correct to four significant figures.

**Order of Convergence:** We consider the order of convergence of the Newton-Raphson method given by the formula,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Let us assume that the sequence of iterations $\{x_n\}$ converges to the root $\xi$. Then, expanding by Taylor's series about $x_n$, the relation $f(\xi) = 0$, gives

$$f(x_n) + (\xi - x_n) f'(x_n) + \frac{1}{2}(\xi - x_n)^2 f''(x_n) + \ldots = 0$$

$$\therefore \quad -\frac{f(x_n)}{f'(x_n)} = \xi - x_n + \frac{1}{2}(\xi - x_n)^2 \cdot \frac{f''(x_n)}{f'(x_n)} + \ldots$$

$$\therefore \quad x_{n+1} - \xi \approx \frac{1}{2}(\xi - x_n)^2 \cdot \frac{f''(x_n)}{f'(x_n)}$$

Taking $\varepsilon_n$ as the error in the nth iteration and writing $\varepsilon_n = \varepsilon_n - \xi$, we have,

$$\varepsilon_{n+1} \approx \frac{1}{2}\varepsilon_n^{\ 2} \ \frac{f''(\xi)}{f'(\xi)} \qquad (2.14)$$

Thus, $\varepsilon_{n+1} = k\varepsilon^2 n$, where $k$ is a constant.

This shows that the order of convergence of Newton-Raphson method is 2. In other words, the Newton-Raphson method has a quadratic rate of convergence.

NOTES

The condition for convergence of Newton-Raphson method can easily be derived by rewriting the Newton-Raphson iterative scheme as $x_{n+1} = \varphi(x_n)$ with $\varphi(x) = x - \dfrac{f(x)}{f'(x)}$

Hence, using the condition for convergence of the linear iteration method, we can write $\varphi'(x) = \dfrac{f(x)\, f''(x)}{[f'(x)]^2}$

Thus, the sufficient condition for the convergence of Newton-Raphson method is,

$$\left|\frac{f(x)\, f''(x)}{[f'(x)]^2}\right| < 1, \text{ in the interval near the root.}$$

i.e.,
$$|f(x)\, f''(x)| < |f'(x)|^2 \qquad (2.15)$$

**Secant Method**

Secant method can be considered as a discretized form of Newton-Raphson method. The iterative formula for this method is obtained from formula of Newton-Raphson method on replacing the derivative $f'(x_0)$ by the gradient of the chord joining two neighbouring points $x_0$ and $x_1$ on the curve $y = f(x)$.

Thus, we have

$$f'(x_0) \approx \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

The iterative formula is given by,

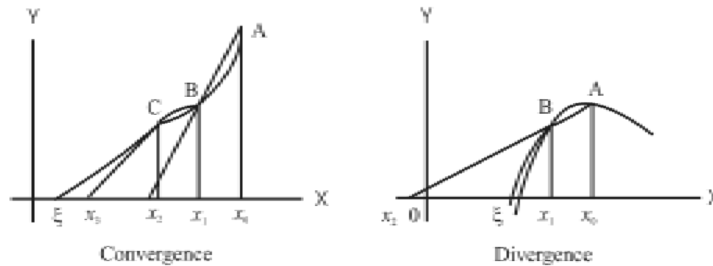
$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

This can be rewritten as,

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)}$$

The iterative formula is equivalent to the one for Regula–Falsi method. The distinction between secant method and Regula–Falsi method lies in the fact that unlike in Regula–Falsi method, the two initial guess values do not bracket a root and the bracketing of the root is not checked during successive iterations, in secant method. Thus, secant method may not always give rise to a convergent sequence to find the root. The geometrical interpretation of the method is shown in Figure 2.4.

NOTES




(Line AB meets $x$-axis alone)

***Fig. 2.4** Secant Method*

**Algorithm:** To find a root of $f(x) = 0$, by Secant method.

**Step 1:** Define $f(x)$.

**Step 2:** Input $x_0, x_1$, *error*, *maxit*. [$x_0, x_1$, are initial guess values, *error* is the prescribed precision and *maxit* is the maximum number of iterations allowed].

**Step 3:** Set $i = 1$

**Step 4:** Compute $f_0 = f(x_0)$

**Step 5:** Compute $f_1 = f(x_1)$

**Step 6:** Compute $x_2 = (x_0 f_1 - x_1 f_0)/(f_1 - f_0)$

**Step 7:** Set $i = i + 1$

**Step 8:** Compute $accy = |x_2 - x_1| / |x_1|$

**Step 9:** Check if $accy < error$, then go to Step 14

**Step 10:** Check if $i \geq maxit$ then go to Step 16

**Step 11:** Set $x_0 = x_1$

**Step 12:** Set $x_1 = x_2$

**Step 13:** Go to step 6

**Step 14:** Print "Root =", $x_2$

**Step 15:** Go to Step 17

**Step 16:** Print 'iterations do not converge'

**Step 17:** Stop

### Regula-Falsi Method

Regula-Falsi method is also a bracketing method. As in bisection method, we start the computation by first finding an interval $(a, b)$ within which a real root lies. Writing $a = x_0$ and $b = x_1$, we compute $f(x_0)$ and $f(x_1)$ and check if $f(x_0)$ and $f(x_1)$ are of opposite signs. For determining the approximate root $x_2$, we find the

NOTES

point of intersection of the chord joining the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$ with the $x$-axis, i.e., the curve $y = f(x_0)$ is replaced by the chord given by,

$$y - f(x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \tag{2.16}$$

Thus, by putting $y = 0$ and $x = x_2$ in Equation (2.16), we get

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0) \tag{2.17}$$

Next, we compute $f(x_2)$ and determine the interval in which the root lies in the following manner. If (*i*) $f(x_2)$ and $f(x_1)$ are of opposite signs, then the root lies in $(x_2, x_1)$. Otherwise if (*ii*) $f(x_0)$ and $f(x_2)$ are of opposite signs, then the root lies in $(x_0, x_2)$. The next approximate root is determined by changing $x_0$ by $x_2$ in the first case and $x_1$ by $x_2$ in the second case.

The aforesaid process is repeated until the root is computed to the desired accuracy $\varepsilon$, i.e., the condition

$|(x_{k+1} - x_k)/x_k| < \varepsilon$, should be satisfied.

Regula-Falsi method can be geometrically interpreted by the following Figure 2.5.

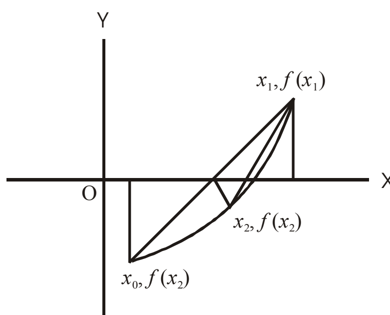



***Fig. 2.5*** *Regula-Falsi Method*

**Algorithm:** Computing root of an equation by Regula-Falsi method.

**Step 1:** Define $f(x)$

**Step 2:** Read epsilon, the desired accuracy

**Step 3:** Read *maxit*, the maximum no. of iterations

**Step 4:** Read $x_0, x_1$ two initial guess values of root

**Step 5:** Compute $f_0 = f(x_0)$

**Step 6:** Compute $f_1 = f(x_1)$

**Step 7:** Check if $f_0 f_1 < 0$, then go to the next step
else go to Step 4

NOTES

**Step 8:** Compute $x_2 = (x_0 f_1 - x_1 f_0) / (f_1 - f_0)$

**Step 9:** Compute $f_2 = f(x_2)$

**Step 10:** Check if $|f_2|$ < epsilon, then go to Step 18

**Step 11:** Check if $f_2 f_0 < 0$ then go to the next Step
else go to Step 15

**Step 12:** Set $x_1 = x_2$

**Step 13:** Set $f_1 = f_2$

**Step 14:** Go to Step 7

**Step 15:** Set $x_0 = x_2$

**Step 16:** Set $f_0 = f_2$

**Step 17:** Go to Step 7

**Step 18:** Write 'root =', $x_2, f_3$

**Step 19:** End

**Example 13:** Use Regula-Falsi method to compute the positive root of $x^3 - 3x - 5 = 0$, correct to four significant figures.

**Solution:** First we find the interval in which the root lies. We observe that $f(2) = -3$ and $f(3) = 13$. Thus, the root lies in [2, 3]. For using the Regula–Falsi method, we use the formula,

$$x_2 = x_0 - \frac{f(x_0)}{f(x_1) - f(x_0)}(x_1 - x_0)$$

With $x_0 = 2$, and $x_1 = 3$, we have

$$x_2 = 2 + \frac{3}{13+3}(3-2) = 2.1875$$

Again, since $f(x_2) = f(2.1875) = -1.095$, we consider the interval [2.1875, 3]. The next approximation is $x_3 = 2.2461$. Also, $f(x_3) = -0.4128$. Hence, the root lies in [2.2461, 3]

Repeating the iterations, we get

$$x_4 = 2.2684, f(x_4) = -0.1328$$
$$x_5 = 2.2748, f(x_5) = -0.0529$$
$$x_6 = 2.2773, f(x_6) = -0.0316$$
$$x_7 = 2.2788, f(x_7) = -0.0028$$
$$x_8 = 2.2792, f(x_8) = -0.0022$$

The root correct to four significant figures is 2.279.

NOTES

**Check Your Progress**

1. How will you compute the roots of the form $f(x) = 0$?
2. Define tabulation method.
3. Explain bisection method.
4. How is order of convergence determined?
5. Explain Newton-Raphson method.
6. Define secant method.
7. Explain Regula-Falsi method.

## 2.3 POLYNOMIAL EQUATIONS

Polynomial equations with real coefficients have some important characteristics regarding their roots. A polynomial equation of degree $n$ is of the form $p_n(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + ... + a_2 x^2 + a_1 x + a_0 = 0$.

(*i*) A polynomial equation of degree $n$ has exactly $n$ roots.

(*ii*) Complex roots occur in pairs, i.e., if $\alpha + i\beta$ is a root of $p_n(x) = 0$, then $\alpha - i\beta$ is also a root.

(*iii*) Descarte's rule of signs can be used to determine the number of possible real roots (positive or negative).

(*iv*) If $x_1, x_2, ..., x_n$ are all real roots of the polynomial equation, then we can express $p_n(x)$ uniquely as,

$$p_n(x) = a_n(x - x_1)(x - x_2)...(x - x_n)$$

(*v*) $p_n(x)$ has a quadratic factor for each pair of complex conjugate roots. Let, $\alpha + i\beta$ and $\alpha - i\beta$ be the roots, then $\{x^2 - 2\alpha x + (\alpha^2 + \beta^2)\}$ is the quadratic factor.

(*vi*) There is a special method, known as Horner's method of synthetic substitution, for evaluating the values of a polynomial and its derivatives for a given $x$.

### Descarte's Rule

The number of positive real roots of a polynomial equation is equal to the number of changes of sign in $p_n(x)$, written with descending powers of $x$, or less by an even number.

Consider for example, the polynomial equation,

$$3x^5 + 2x^4 + x^3 - 2x^2 + x - 2 = 0$$

NOTES

Clearly there are three changes of sign and hence the number of positive real roots is three or one. Thus, it must have a real root. In fact, every polynomial equation of odd degree has a real root.

We can also use Descarte's rule to determine the number of negative roots by finding the number of changes of signs in $p_n(-x)$. For the above equation, $p_n(-x) = -3x^5 + 2x^4 - x^3 - 2x^2 - x - 2 = 0;$ and it has two changes of sign. Thus, it has either two negative real roots or none.

**Check Your Progress**

8. Define polynomial equations.
9. Give the statement of Descarte's rule.

## 2.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. We consider numerical methods for computing the roots of an equation of the form,

$$f(x) = 0$$

   Where $f(x)$ is a reasonably well-behaved function of a real variable $x$.
2. In the tabulation method, a table of values of $f(x)$ is made for values of $x$ in a particular range. Then, we look for the change in sign in the values of $f(x)$ for two consecutive values of $x$. We conclude that a real root lies between these values of $x$.
3. The bisection method involves successive reduction of the interval in which an isolated root of an equation lies.

   The sub-interval in which the root lies is again bisected and the above process is repeated until the length of the sub-interval is less than the desired accuracy.

   The bisection method is also termed as a bracketing method, since the method successively reduces the gap between the two ends of an interval surrounding the real root, i.e., brackets the real root.
4. The order of convergence of an iterative process is determined in terms of the errors $e_n$ and $e_{n+1}$ in successive iterations.
5. Newton-Raphson method is a widely used numerical method for finding a root of an equation $f(x) = 0$, to the desired accuracy. It is an iterative method which has a faster rate of convergence and is very useful when the expression for the derivative $f(x)$ is not complicated. To derive the formula for this method, we consider a Taylor's series expansion of $f(x_0 + h)$, $x_0$ being an initial guess of a root of $f(x) = 0$ and $h$ is a small correction to the root.

6. Secant method can be considered as a discretized form of Newton-Raphson method. The iterative formula for this method is obtained from formula of Newton-Raphson method on replacing the derivative $f'(x_0)$ by the gradient of the chord joining two neighbouring points $x_0$ and $x_1$ on the curve $y = f(x)$.
7. Regula-Falsi method is also a bracketing method. As in bisection method, we start the computation by first finding an interval $(a, b)$ within which a real root lies. Writing $a = x_0$ and $b = x_1$, we compute $f(x_0)$ and $f(x_1)$ and check if $f(x_0)$ and $f(x_1)$ are of opposite signs. For determining the approximate root $x_2$, we find the point of intersection of the chord joining the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$ with the $x$-axis, i.e., the curve $y = f(x_0)$ is replaced by the chord given by,

$$y - f(x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

8. A polynomial equation of degree $n$ is of the form $p_n(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + ... + a_2 x^2 + a_1 x + a_0 = 0$.
9. The number of positive real roots of a polynomial equation is equal to the number of changes of sign in $p_n(x)$, written with descending powers of $x$, or less by an even number.

## 2.5 SUMMARY

- A root of an equation is usually computed in two stages. First, we find the location of a root in the form of a crude approximation of the root. Next we use an iterative technique for computing a better value of the root to a desired accuracy in successive approximations/computations.
- Tabulation Method: In the tabulation method, a table of values of $f(x)$ is made for values of $x$ in a particular range.
- The bisection method involves successive reduction of the interval in which an isolated root of an equation lies.
- If a function $f(x)$ is continuous in the closed interval $[a, b]$ and $f(a)$ and $f(b)$ are of opposite signs, i.e., $f(a) f(b) < 0$, then there exists at least one real root of $f(x) = 0$ between $a$ and $b$.
- The bisection method is also termed as a bracketing method, since the method successively reduces the gap between the two ends of an interval surrounding the real root, i.e., brackets the real root.
- If the function $g(x)$ is continuous in the interval $[a, b]$ which contains a root $\xi$ of the equation $f(x) = 0$, and is rewritten as $x = g(x)$, and $|g'(x)| \le l \le 1$ in this interval, then for any choice of $x_0 \in [a, b]$, the sequence $\{x_n\}$ determined by the iterations,

$x_{k+1} = g(x_k)$, for $k = 0, 1, 2, \ldots$

This converges to the root of $f(x) = 0$.

- Order of Convergence: The order of convergence of an iterative process is determined in terms of the errors $e_n$ and $e_{n+1}$ in successive iterations. An iterative process is said to have $k$th order convergence if $\lim_{n\to\infty} \frac{e_{n+1}}{e_n^k} < M$, where $M$ is a finite number.
- Newton-Raphson method is a widely used numerical method for finding a root of an equation $f(x) = 0$, to the desired accuracy.
- Secant method can be considered as a discretized form of Newton-Raphson method. The iterative formula for this method is obtained from formula of Newton-Raphson method on replacing the derivative $f'(x_0)$ by the gradient of the chord joining two neighbouring points $x_0$ and $x_1$ on the curve $y = f(x)$.
- Descarte's rule of signs can be used to determine the number of possible real roots (positive or negative).
- If $x_1, x_2, \ldots, x_n$ are all real roots of the polynomial equation, then we can express $p_n(x)$ uniquely as,

  $p_n(x) = a_n(x - x_1)(x - x_2)\ldots(x - x_n)$
- We can also use Descarte's rule to determine the number of negative roots by finding the number of changes of signs in $p_n(-x)$.

## 2.6 KEY WORDS

- **Graphical Method:** In the graphical method, we draw the graph of the function $y = f(x)$, for a certain range of values of $x$.
- **Tabulation Method:** In the tabulation method, a table of values of $f(x)$ is made for values of $x$ in a particular range.
- **Order of Convergence:** The order of convergence of an iterative process is determined in terms of the errors $e_n$ and $e_{n+1}$ in successive iterations.

## 2.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. What is tabulation method?
2. What is bisection method?

3. Define Newton-Rapshon method.
4. What is meant by secant method?
5. Explain Regula-Falsi method.

NOTES

**Long-Answer Questions**

1. Use graphical method to find the location of a real root of the equation $x^3 + 10x - 15 = 0$.
2. Draw the graphs of the function $f(x) = \cos x - x$, in the range $[0, \pi/2)$ and find the location of the root of the equation $f(x) = 0$.
3. Compute the root of the equation $x^3 - 9x + 1 = 0$ which lies between 2 and 3 correct upto three significant digits using bisection method.
4. Compute the root of the equation $x^3 + x^2 - 1 = 0$, near 1, by the iterative method correct upto two significant digits.
5. Use iterative method to find the root near $x = 3.8$ of the equation $2x - \log_{10}x = 7$ correct upto four significant digits.
6. Compute using Newton-Raphson method the root of the equation $e^x = 4^x$, near 2, correct upto four significant digits.
7. Use an iterative formula to compute $\sqrt[7]{125}$ correct upto four significant digits.
8. Find the real root of $x \log_{10}x - 1.2 = 0$ correct upto four decimal places using Regula-Falsi method.
9. Use Regula-Falsi method to find the root of the following equations correct upto four significant figures:
   (*i*) $x^3 - 4x - 1 = 0$, the root near $x = 2$
   (*ii*) $x^6 - x^4 - x^3 - 1 = 0$, the root between 1.4 and 1.5
10. Compute the positive root of the given equation correct upto four places of decimals using Newton-Raphson method:

$$x + \log_e x = 2$$

## 2.8 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

NOTES

# UNIT 3 BIRGE – VIETA, BAIRSTOW'S AND GRAEFFE'S ROOT SQUARING METHODS

**Structure**



## 3.0 INTRODUCTION

In mathematics, a polynomial is an expression consisting of variables (also called indeterminate) and coefficients that involves only the operations of addition, subtraction, multiplication, and non-negative integer exponents of variables. The polynomial of a single indeterminate, $x$, is $x^2 - 4x + 7$ while in three variables it is $x^3 + 2xyz^2 - yz + 1$. A polynomial equation is, therefore, an equation that has multiple terms made up of numbers and variables. The degree tells us how many roots can be found in a polynomial equation. For example, if the highest exponent is 3, then the equation has three roots. The roots of the polynomial equation are the values of $x$ where $y = 0$. Principally, the polynomial equation is an equation of the form $f(x) = 0$ where $f(x)$ is a polynomial in $x$.

The sixteenth century French mathematician Francois Vieta was the pioneer to develop methods for finding approximate roots of polynomial equations. Later, several other methods were developed for solving polynomial equations. In numerical analysis, Bairstow's method is an efficient algorithm for finding the roots of a real polynomial of arbitrary degree. The algorithm first appeared in the appendix of the 1920 book '*Applied Aerodynamics*' by Leonard Bairstow. The algorithm finds the roots in complex conjugate pairs using only real arithmetic. The Graeffe's root squaring method is a direct method to find the roots of any polynomial equation with real coefficients. Polynomials are used to form polynomial equations, which encode a wide range of problems, from elementary word problems to complicated scientific problems.

In this unit, you will study about the Birge-Vieta method, the Bairstow's method, and the Graeffe's root squaring method.

## 3.1 OBJECTIVES

NOTES

After going through this unit, you will be able to:

- Discuss the Birge-Vieta method
- Understand the Bairstow's method
- Elaborate on the Graeffe's root squaring method

## 3.2 BIRGE – VIETA METHOD

Birge-Vieta method is used for finding the real roots of a polynomial equations. This method is based on an original method developed by the two English mathematicians Birge and Vieta. Finding and approximating the derivation of all roots of a polynomial equation is a very significant. In the field of science and engineering, there are numerous applications which require the solutions of all roots of a polynomial equations for a particular problem.

Newton-Raphson method is fundamentally used for finding the root of an algebraic and transcendental equations. Since the rate of convergence of this method is quadratic, hence the Newton-Raphson method can be used to find a root of a polynomial equation as polynomial equation is an algebraic equation. Birge-Vieta method is based on the Newton-Raphson method or this method is a modified form of Newton-Raphson method.

Consider the given polynomial equation of degree $n$, which has the form,

$P_n(x) = a_n x^n + \dots + a_1 x + a_0 = 0.$

Let $x_0$ be an initial approximation to the root $\alpha$. The Newton-Raphson iterated formula for improving this approximation is,

$$x_i = x_{i-1} - \frac{p_n(x_{i-1})}{p'_n(x_{i-1})}, \quad i = 1, 2, \dots$$

To apply this formula, first evaluate both $P_n(x)$ and $P'_n(x_i)$ at any $x_i$. The utmost natural method is to evaluate,

$$p_n(x_i) = a_n x_i^n + a_{n-1} x_i^{n-1} + \dots + a_2 x_i^2 + a_1 x_i + a_0$$

$$p'_n(x_i) = n\, a_n x_i^{n-1} + (n-1)\, a_{n-1} x_i^{n-2} + \dots + 2a_2 x_i + a_1$$

However, this is stated as the most inefficient method of evaluating a polynomial, because of the amount of computations involved and also because of the possible growth of round off errors. Thus there must be some proficient and effective method for evaluating $P_n(x)$ and $P'_n(x)$.

Vieta's formula is used for the coefficients of polynomial to the sum and product of their roots, along with the products of the roots that are in groups. Vieta's formula defines the association of the roots of a polynomial by means of its coefficients. Following example will make the concept clear that how to find a polynomial with given roots.

Here we will discuss about the real-valued polynomials, i.e., the coefficients of polynomials are real numbers.

Consider a quadratic polynomial. If the given two real roots are $r_1$ and $r_2$, then find a polynomial.

Let the polynomial is $a_2x^2 + a_1x + a_0$. When the roots are given, then we can also write the polynomial equation in the form, $k(x - r_1)(x - r_2)$.

Since both the equations denotes the same polynomial, therefore equate both polynomials as,

$$a_2x^2 + a_1x + a_0 = k(x - r_1)(x - r_2) \qquad (3.1)$$

On simplifying the Equation (3.1), we have the following form of equation,

$$a_2x^2 + a_1x + a_0 = kx^2 - k(r_1 + r_2)x + k(r_1r_2)$$

Comparing the coefficients of both the sides of the above equation, we have,

For $x^2$, $a_2 = k$

For $x$, $a_1 = -k(r_1 + r_2)$

For constant term, $a_0 = k\,r_1r_2$

Which gives,

$$a_2 = k$$

Therefore,

$$\frac{a_1}{a_2} = -(r_1 + r_2) \qquad (3.2)$$

$$\frac{a_0}{a_2} = r_1r_2 \qquad (3.3)$$

Equations (3.2) and (3.4) are termed as Vieta's formulas for a second degree polynomial.

As a general rule, for an $n$th degree polynomial, there are $n$ different Vieta's formulas which can be written in a condensed form as,

**NOTES**

For $0 \leq k \leq n$

$$\sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} (r_{i_1} r_{i_2} \ldots r_{i_k}) = (-1)^k \frac{a_{n-k}}{a_n}$$

**Example 1:** Find all the roots of the given polynomial equation $P_3(x) = x^3 + x - 3 = 0$ rounded off to three decimal places. Stop the iteration whenever $\{x_{i+1} - x_i\} < 0.0001$.

**Solution:** The equation $P_3(x) = 0$ has three roots. Since there is only one change in the sign of the coefficients, the equation can have as a maximum one positive real root. The equation has no negative real root since $P_3(-x) = 0$ has no change of sign of coefficients. Since $P_3(x) = 0$ is of odd degree it has at least one real root. Hence the given equation $x^3 + x - 3 = 0$ has one positive real root and a complex pair. Since P(1) = –1 and P(2) = 7, as per the intermediate value theorem the equation has a real root lying in the interval ]1,2[.

Now we will find the real root using Birge-Vieta method. Let the initial approximation be 1.1.

**First Iteration**

| | | | | |
|---|---|---|---|---|
| | 1 | 0 | 14 | –3 |
| 1.1 | | 1.1 | 1.21 | 2.431 |
| | 1 | 1.1 | 2.21 | –0.569 |
| 1.1 | | 1.1 | 2.42 | |
| | 1 | 2.2 | 4.63 | |

Therefore $x_1 = 1.1 - (-0.569)/4.63 = 1.22289$

Similarly,

$x_2 = 1.21347$

$x_3 = 1.21341$

Since, $|x_2 - x_3| < 0.0001$, we stop the iteration here. Hence the required value of the root is 1.213, rounded off to three decimal places.

Next we will find the deflated polynomial of $P_3(x)$. To obtain the deflated polynomial, we have to first find the polynomial $q_2(x)$ by using the final approximation $x_3 = 1.213$, as shown in the following table.

| | | | | |
|---|---|---|---|---|
| | 1 | 0 | 1 | –3 |
| 1.213 | | 1.213 | 1.4714 | 2.9978 |
| | 1 | 1.213 | 2.4714 | –0.0022 |

Here, $P_3(1.213) = -0.0022$, i.e., the magnitude of the error in satisfying $P_3(x_3) = 0$ is 0.0022.

We then find $q_2(x) = x^2 + 1.213\,x + 2.4714 = 0$

NOTES

This is a quadratic equation and its roots are given by,

$$x = \frac{-1.213 \pm \sqrt{(1.213)^2 - 4 \times 2.4714}}{2}$$

$$= \frac{-1.213 \pm 2.9009\, i}{2}$$

$$= 0.6065 \ \pm 1.4505\ i$$

Hence the three roots of the equation rounded off to three decimal places are 1.213, 0.6065 + 1.4505 i and – 0.6065 – 1.4505 i.

## 3.3 BAIRSTOW'S METHOD

In numerical analysis, Bairstow's method is an efficient algorithm for finding the roots of a real polynomial of arbitrary degree. The algorithm was formulated by Leonard Bairstow and first appeared in the appendix of the book '*Applied Aerodynamics*' (1920). The algorithm finds the roots in complex conjugate pairs using only real arithmetic.

Bairstow's approach is to use Newton's method to adjust the coefficients $u$ and $v$ in the quadratic $x^2 + ux + v$ until its roots are also roots of the polynomial being solved. The roots of the quadratic may then be determined, and the polynomial may be divided by the quadratic to eliminate those roots. This process is then iterated until the polynomial becomes quadratic or linear, and all the roots have been determined.

Long division of the polynomial to be solved as,

$$P(x) = \sum_{i=0}^{n} a_i x^i$$

By $x^2 + ux + v$ yields a quotient,

$$Q(x) = \sum_{i=0}^{n-2} b_i x^i$$

And a remainder $cx + d$ such that,

$$P(x) = (x^2 + ux + v)\left(\sum_{i=0}^{n-2} b_i x^i\right) + (cx + d)$$

A second division of $Q(x)$ by $x^2 + ux + v$ is performed to yield a quotient,

$$R(x) = \sum_{i=0}^{n-4} f_i x^i$$

And a remainder $gx + h$ with,

$$Q(x) = (x^2 + ux + v)\left(\sum_{i=0}^{n-4} f_i x^i\right) + (gx + h)$$

The variables $c, d, g, h$ and the $\{b_i\}$, $\{f_i\}$ are functions of $u$ and $v$. They can be found recursively as follows,

$$\begin{aligned} b_n &= b_{n-1} = 0, & f_n &= f_{n-1} = 0, & \\ b_i &= a_{i+2} - ub_{i+1} - vb_{i+2} & f_i &= b_{i+2} - uf_{i+1} - vf_{i+2} & (i = n-2, \ldots, 0), \\ c &= a_1 - ub_0 - vb_1, & g &= b_1 - uf_0 - vf_1, & \\ d &= a_0 - vb_0, & h &= b_0 - vf_0. & \end{aligned}$$

The quadratic evenly divides the polynomial when,

$c(u, v) = d(u, v) = 0$

Values of $u$ and $v$ for which this occurs can be discovered by picking starting values and iterating Newton's method in two dimensions as,

$$\begin{bmatrix} u \\ v \end{bmatrix} := \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} \frac{\partial c}{\partial u} & \frac{\partial c}{\partial v} \\ \frac{\partial d}{\partial u} & \frac{\partial d}{\partial v} \end{bmatrix}^{-1} \begin{bmatrix} c \\ d \end{bmatrix} := \begin{bmatrix} u \\ v \end{bmatrix} - \frac{1}{vg^2 + h(h - ug)} \begin{bmatrix} -h & g \\ -gv & gu - h \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix}$$

This continues until convergence occurs. This method to find the zeroes of polynomials can thus be easily implemented with a programming language or even a spreadsheet.

**Example 2:** The task is to determine a pair of roots of the polynomial,

$f(x) = 6x^5 + 11x^4 - 33x^3 - 33x^2 + 11x + 6$

**Solution:** As first quadratic polynomial we can use the normalized polynomial formed from the leading three coefficients of $f(x)$,

$$u = \frac{a_{n-1}}{a_n} = \frac{11}{6}; \quad v = \frac{a_{n-2}}{a_n} = -\frac{33}{6}$$

The iteration then produces as shown in the following table.

### Iteration Steps of Bairstow's Method

| Nr | u | v | step length | roots |
|---|---|---|---|---|
| 0 | 1.833333333333 | −5.500000000000 | 5.579008780071 | −0.916666666667±2.517990821623 |
| 1 | 2.979026068546 | −0.039896784438 | 2.048558558641 | −1.489513034273±1.502845921479 |
| 2 | 3.635306053091 | 1.900693009946 | 1.799922838287 | −1.817653026545±1.184554563945 |
| 3 | 3.064938039761 | 0.193530875538 | 1.256481376254 | −1.532469019881±1.467968126819 |
| 4 | 3.461834191232 | 1.385679731101 | 0.428931413521 | −1.730917095616±1.269013105052 |
| 5 | 3.326244386565 | 0.978742927192 | 0.022431883898 | −1.663122193282±1.336874153612 |
| 6 | 3.333340909351 | 1.000022701147 | 0.000023931927 | −1.666670454676±1.333329555414 |
| 7 | 3.333333333340 | 1.000000000020 | 0.000000000021 | −1.666666666670±1.333333333330 |
| 8 | 3.333333333333 | 1.000000000000 | 0.000000000000 | −1.666666666667±1.333333333333 |

After eight iterations the method produced a quadratic factor that contains the roots –1/3 and –3 within the represented precision. The step length from the fourth iteration on demonstrates the superlinear speed of convergence.



## 3.4 GRAEFFE'S ROOT SQUARING METHOD

In mathematics, Graeffe's method or Dandelin–Lobachesky–Graeffe method is an algorithm typically used for finding all of the roots of a polynomial. It was developed independently by Germinal Pierre Dandelin in 1826 and Lobachevsky in 1834. In 1837 Karl Heinrich Gräffe also discovered the principal idea of the method. The method separates the roots of a polynomial by squaring them repeatedly. This squaring of the roots is done implicitly, that is, only working on the coefficients of the polynomial. Finally, Viète's formulas are used in order to approximate the roots.

### Dandelin–Graeffe Iteration

Let $p(x)$ be a polynomial of degree $n$,

$$p(x) = (x - x_1) \cdots (x - x_n)$$

Then,

$$p(-x) = (-1)^n (x + x_1) \cdots (x + x_n)$$

Let $q(x)$ be the polynomial which has the squares $x_1^2, \cdots, x_n^2$ as its roots,

$$q(x) = \left(x - x_1^2\right) \cdots \left(x - x_n^2\right)$$

Then we can denote as,

$$\begin{aligned}
q(x^2) &= \left(x^2 - x_1^2\right) \cdots \left(x^2 - x_n^2\right) \\
&= (x - x_1)(x + x_1) \cdots (x - x_n)(x + x_n) \\
&= \{(x - x_1) \cdots (x - x_n)\} \times \{(x + x_1) \cdots (x + x_n)\} \\
&= p(x) \times \{(-1)^n (-x - x_1) \cdots (-x - x_n)\} \\
&= p(x) \times \{(-1)^n p(-x)\} \\
&= (-1)^n p(x) p(-x)
\end{aligned}$$

Next $q(x)$ can now be computed by algebraic operations on the coefficients of the polynomial $p(x)$ alone.

Let,

$$\begin{aligned}
p(x) &= x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n \\
q(x) &= x^n + b_1 x^{n-1} + \cdots + b_{n-1} x + b_n
\end{aligned}$$

Then the coefficients are related by,

$$b_k = (-1)^k a_k^2 + 2\sum_{j=0}^{k-1}(-1)^j\, a_j a_{2k-j}, \qquad a_0 = b_0 = 1$$

Graeffe observed that if one separates $p(x)$ into its odd and even parts, then

$$p(x) = p_e\left(x^2\right) + xp_o\left(x^2\right)$$

We now obtain a simplified algebraic expression for $q(x)$ of the form,

$$q(x) = (-1)^n\left(p_e(x)^2 - xp_o(x)^2\right)$$

This expression involves the squaring of two polynomials of only half the degree, and is therefore used in most implementations of the method.

Iterating this procedure several times separates the roots with respect to their magnitudes. Repeating $k$ times gives a polynomial of degree $n$, we have:

$$q^k(y) = y^n + a^k{}_1\, y^{n-1} + \cdots + a^k{}_{n-1}\, y + a^k{}_n$$

With roots,

$$y_1 = x_1^{2^k},\ y_2 = x_2^{2^k},\ \ldots,\ y_n = x_n^{2^k}.$$

If the magnitudes of the roots of the original polynomial were separated by some factor $\rho > 1$, that is, $|x_k| \geq \rho|x_{k+1}|$, then the roots of the $k$-th iterate are separated by a fast growing factor,

$$\rho^{2^k} \geq 1 + 2^k(\rho - 1)$$

Next the Vieta relations are used as Classical Graeffe's method as shown below:

$$\begin{aligned}
a^k{}_1 &= -(y_1 + y_2 + \cdots + y_n)\\
a^k{}_2 &= y_1y_2 + y_1y_3 + \cdots + y_{n-1}y_n\\
&\ \vdots\\
a^k{}_n &= (-1)^n(y_1y_2\cdots y_n).
\end{aligned}$$

If the roots $x_1, \ldots, x_n$ are sufficiently separated, say by a factor $\rho > 1$, $|x_m| \geq \rho|x_{m+1}|$, then the iterated powers $y_1, y_2, \ldots, y_n$ of the

roots are separated by the factor $\rho^{2^k}$, which quickly becomes very big. The coefficients of the iterated polynomial can then be approximated by their leading term,

$$a^k_1 \approx -y_1$$

$$a^k_2 \approx y_1 y_2 \text{ and so on,}$$

Implying,

$$y_1 \approx -a^k_1,\ y_2 \approx -a^k_2/a^k_1,\ \ldots\ y_n \approx -a^k_n/a^k_{n-1}.$$

Finally, logarithms are used in order to find the absolute values of the roots of the original polynomial. These magnitudes alone are already useful to generate meaningful starting points for other root-finding methods.

**Check Your Progress**

1. Why the Birge-Vieta method is used?
2. How the Bairstow's approach uses Newton's method for adjusting the coefficients u and v in the quadratic $x^2 + ux + v$?
3. Explain Graeffe's method.

## 3.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Birge-Vieta method is used for finding the real roots of a polynomial equations.
2. Bairstow's approach is to use Newton's method to adjust the coefficients $u$ and $v$ in the quadratic $x^2 + ux + v$ until its roots are also roots of the polynomial being solved. The roots of the quadratic may then be determined, and the polynomial may be divided by the quadratic to eliminate those roots. This process is then iterated until the polynomial becomes quadratic or linear, and all the roots have been determined.
3. Graeffe's method or Dandelin–Lobachesky–Graeffe method is an algorithm typically used for finding all of the roots of a polynomial. It was developed independently by Germinal Pierre Dandelin in 1826 and Lobachevsky in 1834. The method separates the roots of a polynomial by squaring them repeatedly. This squaring of the roots is done implicitly, that is, only working on the coefficients of the polynomial. Finally, Viète's formulas are used in order to approximate the roots.

## 3.6 SUMMARY

NOTES

- Birge-Vieta method is used for finding the real roots of a polynomial equations. This method is based on an original method developed by the two English mathematicians Birge and Vieta.
- Finding and approximating the derivation of all roots of a polynomial equation is a very significant. In the field of science and engineering, there are numerous applications which require the solutions of all roots of a polynomial equations for a particular problem.
- Newton-Raphson method is fundamentally used for finding the root of an algebraic and transcendental equations.
- Since the rate of convergence of this method is quadratic, hence the Newton-Raphson method can be used to find a root of a polynomial equation as polynomial equation is an algebraic equation.
- Birge-Vieta method is based on the Newton-Raphson method or this method is a modified form of Newton-Raphson method.
- The most inefficient method of evaluating a polynomial, because of the amount of computations involved and also because of the possible growth of round off errors. Thus there must be some proficient and effective method for evaluating $P_n(x)$ and $P'_n(x)$.
- Vieta's formula is used for the coefficients of polynomial to the sum and product of their roots, along with the products of the roots that are in groups.
- Vieta's formula defines the association of the roots of a polynomial by means of its coefficients. Following example will make the concept clear that how to find a polynomial with given roots.
- As a general rule, for an $n$th degree polynomial, there are $n$ different Vieta's formulas which can be written in a condensed form as,

  $0 \leq k \leq n$

  For

  $$\sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} (r_{i_1} r_{i_2} \ldots r_{i_k}) = (-1)^k \frac{a_{n-k}}{a_n}$$

- In numerical analysis, Bairstow's method is an efficient algorithm for finding the roots of a real polynomial of arbitrary degree.
- The algorithm was formulated by Leonard Bairstow. The algorithm finds the roots in complex conjugate pairs using only real arithmetic.
- Bairstow's approach uses Newton's method to adjust the coefficients $u$ and $v$ in the quadratic $x^2 + ux + v$ until its roots are also roots of the polynomial being solved.

NOTES

- The roots of the quadratic may then be determined, and the polynomial may be divided by the quadratic to eliminate those roots. This process is then iterated until the polynomial becomes quadratic or linear, and all the roots have been determined.
- In mathematics, Graeffe's method or Dandelin–Lobachesky–Graeffe method is an algorithm typically used for finding all of the roots of a polynomial. It was developed independently by Germinal Pierre Dandelin in 1826 and Lobachevsky in 1834. In 1837 Karl Heinrich Gräffe also discovered the principal idea of the method.
- The Graeffe's method separates the roots of a polynomial by squaring them repeatedly. This squaring of the roots is done implicitly, that is, only working on the coefficients of the polynomial. Finally, Viète's formulas are used in order to approximate the roots.
- In Graeffe's method, the logarithms are used in order to find the absolute values of the roots of the original polynomial. These magnitudes alone are already useful to generate meaningful starting points for other root-finding methods.

## 3.7 KEY WORDS

- **Birae-Vieta method:** This method is used for finding the real roots of a polynomial equations.
- **Bairstow's method:** This is an efficient algorithm for finding the roots of a real polynomial of arbitrary degree. The algorithm was formulated by Leonard Bairstow for finding the roots in complex conjugate pairs using only real arithmetic.
- **Graeffe's method or Dandelin–Lobachesky–Graeffe method:** It is an algorithm typically used for finding all of the roots of a polynomial.

## 3.8 SELF-ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. Why is Birge-Vieta method used?
2. Define Bairstow's method.
3. What is the significance of Graeffe's root squaring method?

### Long-Answer Questions

1. Briefly explain the Birge-Vieta method giving appropriate examples.
2. Find the root of $x^4 - 3x^3 + 3x^2 - 3x + 2 = 0$ using Birge-Vieta method.

3. Explain the Bairstow's method with the help of examples.
4. Using Bairstow's method find all the roots of a given polynomial,

$$f_n(x) = a_0 + a_1x + a_2x^2 + ... + a_nx^n$$

5. Discuss the Graeffe's root squaring method giving appropriate examples.

## 3.9 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

# UNIT 4 SOLUTION OF SIMULTANEOUS LINEAR EQUATION

**Structure**



## 4.0 INTRODUCTION

Many engineering and scientific problems require the solution based on system of linear equations. The system of equations is termed as a homogeneous type if all the elements in the column vector $b$ are zero else the system is termed as a non-homogeneous type. You will learn the method of computation to find the solution of a system of $n$ linear equations in $n$ unknowns. Two types of efficient numerical methods are used for computing solution of systems of equations, of which some are direct methods and others are iterative in nature. In the direct method, Gaussian elimination method is used while in the iterative method, Gauss-Seidel iteration method is commonly used. You will learn the two forms of iteration methods termed as Jacobi iteration method and Gauss-Seidel iteration method.

In this unit, you will study about the transcendental and polynomial equations and rate of convergence of iterative methods.

## 4.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the system of linear equations
- Understand Cramer's rule
- Explain Gaussian elimination method and Gauss-Jordan elimination method

- Define Jacobi iteration method and Gauss-Seidel iteration method
- Compute inverse of a matrix using Gaussian elimination method

NOTES

## 4.2 SYSTEM OF LINEAR EQUATIONS

Many engineering and scientific problems require the solution of a system of linear equations. We consider a system of $m$ linear equations in $n$ unknowns written as,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + ... + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + ... + a_{3n}x_n &= b_3 \\ ... \quad ... \quad ... & \\ a_{m1} + a_{m2}x_2 + a_{m3}x_3 + ... + a_m x_n &= b_m \end{aligned} \qquad (4.1)$$

Using matrix notation, we can write the above system of equations in the form,

$$Ax = b \qquad (4.2)$$

where $A$ is a $m \times n$ matrix and $x$, $b$ are respectively $n$-column, $m$-row vectors given by,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ a_{31} & a_{32} & a_{33} & ... & a_{3n} \\ ... & ... & ... & ... & ... \\ a_{m1} & a_{m2} & a_{m3} & ... & a_{mn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ ... \\ b_m \end{bmatrix} \qquad (4.3)$$

The system of equations is termed as a homogeneous one, if all the elements in the column vector $b$ are zero. Otherwise, the system is termed as a non-homogeneous one.

The homogeneous system has a non-trivial solution, if $A$ is a square matrix, i.e., $m = n$, and the determinant of the coefficient matrix, i.e., $|A|$ is equal to zero.

The solution of the non-homogeneous system exists, if the rank of the coefficient matrix $A$ is equal to the rank of the augmented matrix $[A : b]$ given by,

$$[A : b] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & ... & a_{3n} & b_3 \\ ... & ... & ... & ... & ... & ... \\ a_{n1} & a_{n2} & an_3 & ... & a_{nn} & b_n \end{bmatrix}$$

Further, a unique non-trivial solution of the system given by Equation (4.1) exists when $m = n$ and the determinant $|A| \neq 0$, i.e., the coefficient matrix is a square non-singular matrix. The computation of the solution of a system of $n$ linear equations in $n$ unknowns can be made by any one of the two classical methods

NOTES

known as the Cramer's rule and the matrix inversion method. But these two methods are not suitable for numerical computation, since both the methods require the evaluation of determinants. There are two types of efficient numerical methods for computing solution of systems of equations. Some are direct methods and others are iterative in nature. Among the direct methods, Gaussian elimination method is most commonly used. Among the iterative methods, Gauss-Seidel iteration method is very commonly used.

### 4.2.1 Classical Methods

**Cramer's Rule:** Let $D = |A|$ be the determinant of the coefficient matrix $A$ and $D_i$ be the determinant obtained by replacing the $i$th column of $D$ by the column vector $b$. The Cramer's rule gives the solution vector $x$ by the equations,

$$x_i = \frac{D_i}{D}, \text{ for } i = 1, 2, ..., n \tag{4.4}$$

Thus we have to compute $(n + 1)$ determinants of order $n$.

**Example 1:** Use Cramer's rule to solve the following system:

$$\begin{bmatrix} 2 & -3 & 1 \\ 3 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

**Solution:** The determinant $D$ of the coefficient matrix is,

$$D = \begin{bmatrix} 2 & -3 & 1 \\ 3 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix} = 2(-1-1) - 3(-1+3) + (-3-1) = -14$$

The determinants $D_1$, $D_2$ and $D_3$ are,

$$D_1 = \begin{vmatrix} 1 & -3 & 1 \\ 2 & 1 & -1 \\ 1 & -1 & -1 \end{vmatrix} = (-1-1) - 3(-1+2) + (-2-1) = -8$$

$$D_2 = \begin{vmatrix} 2 & 1 & 1 \\ 3 & 2 & -1 \\ 1 & 1 & -1 \end{vmatrix} = 2(-2+1) + (-1+3) + (3-2) = 1$$

$$D_3 = \begin{vmatrix} 2 & -3 & 1 \\ 3 & 1 & 2 \\ 1 & -1 & 1 \end{vmatrix} = 2(1+2) - 3(2-3) + (-3-1) = 5$$

Hence by Cramer's rule, we get

$$x_1 = \frac{D_1}{D} = \frac{-8}{-14} = \frac{4}{7}, \quad x_2 = \frac{D_2}{D} = \frac{-1}{14}, \quad x_3 = \frac{D_3}{D} = -\frac{5}{14}$$

### 4.2.2 Elimination Methods

**Matrix Inversion Method:** Let $A^{-1}$ be the inverse of the matrix $A$ defined by,

$$A^{-1} = \frac{\text{Adj } A}{|A|} \tag{4.5}$$

where Adj $A$ is the adjoint matrix obtained by transposing the matrix of the cofactors of the elements $a_{ij}$ of the determinant of the coefficient matrix $A$.

Thus,

$$\text{Adj } A = \begin{bmatrix} A_{11} & A_{21} \ldots\ldots & A_{n1} \\ A_{12} & A_{22} \ldots\ldots & A_{n2} \\ \ldots & \ldots\ldots & \ldots\ldots \\ A_{1n} & A_{2n} \ldots\ldots & A_{nn} \end{bmatrix} \tag{4.6}$$

$A_{ij}$ being the cofactor of $a_{ij}$.

Then the solution of the system is given by,

$$x = A^{-1}b \tag{4.7}$$

***Note:*** If the rank of the coefficient matrix of a system of linear equations in $n$ unknowns is less than $n$, then there are more unknowns than the number of independent equations. In such a case, the system has an infinite set of solutions.

**Example 2:** Solve the given system of equations by matrix inversion method:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & 3 \\ 3 & 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ 1 \end{bmatrix}$$

**Solution:** For solving the system of equations by matrix inversion method we first compute the determinant of the coefficient matrix,

$$|A| = \begin{vmatrix} 1 & 1 & 1 \\ 2 & -1 & 3 \\ 3 & 2 & -1 \end{vmatrix} = 13$$

Since $|A| \neq 0$, the matrix $A$ is non-singular and $A^{-1}$ exists. We now compute the adjoint matrix,

$$\text{Adj } A = \begin{bmatrix} -5 & 3 & 4 \\ 11 & -4 & -1 \\ 7 & 1 & -3 \end{bmatrix}. \quad \text{Thus,} \quad A^{-1} = \frac{\text{Adj } A}{|A|} = \frac{1}{13}\begin{bmatrix} -5 & 3 & 4 \\ 11 & -4 & -1 \\ 7 & 1 & -3 \end{bmatrix}$$

Hence, the solution by matrix inversion method gives,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = A^{-1}b = \frac{1}{13}\begin{bmatrix} -5 & 3 & 4 \\ 11 & -4 & -1 \\ 7 & 1 & -3 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{13}\begin{bmatrix} -13 \\ 39 \\ 26 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 2 \end{bmatrix}$$

**Gaussian Elimination Method:** This method consists in systematic elimination of the unknowns so as to reduce the coefficient matrix into an upper triangular system, which is then solved by the procedure of back-substitution. To understand the procedure for a system of three equations in three unknowns, consider the following system of equations:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \qquad (4.8(a))$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \qquad (4.8(b))$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \qquad (4.8(c))$$

We have to first eliminate $x_1$ from the last two equations and then eliminate $x_2$ from the last equation.

In order to eliminate $x_1$ from the second equation we multiply the Equation (4.8(a)) by $-a_{21}/a_{11} = m_2$, and add to the second equation. Similarly, for elimination of $x_1$ from the third Equation (4.8(c)) we have to multiply the first Equation (4.8(a)) by $-a_{31}/a_{11} = m_3$, and add to the last Equation (4.8(c)). We would then have the following two equations from them:

$$a_{22}^{(1)}\ x_2 + a_{23}^{(1)}\ x_3 = b_2^{(1)} \qquad (4.9(a))$$

$$a_{32}^{(1)}\ x_2 + a_{33}^{(1)}\ x_3 = b_3^{(1)} \qquad (4.9(b))$$

where $a_{22}^{(1)} = a_{22} - m_2 a_{12},\ \ a_{23}^{(1)} = a_{23} - m_2 a_{13},\ \ b_2^{(1)} = b_2 - m_2 b_1$

$a_{32}^{(1)} = a_{32} - m_2 a_{12},\ \ a_{33}^{(1)} = a_{33} - m_2 a_{13},\ \ b_3^{(1)} = b_3 - m_3 b_1$

Again for eliminating $x_2$ from the last of the above two equations, we multiply the first Equation (4.9(a)) by $m_4 = -a_{32}^{(1)}/a_{22}^{(1)}$, and add to the second Equation (4.9(b)), which would give the equation,

$$a_{33}^{(2)} = b_3^{(2)} \qquad (4.10)$$

where $a_{33}^{(2)} = a_{33}^{(1)} - m_4 a_{23}^{(1)},\ b_3^{(2)} = b_3^{(1)} - m_4 b_2^{(1)}$

Thus by systematic elimination we get the triangular system given below,

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \qquad (4.11(a))$$

$$a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 = b_2^{(1)} \qquad (4.11(b))$$

$$a_{33}^{(2)}x_3 = b_2^{(2)} \qquad (4.11(c))$$

It is now easy to solve the unknowns by back-substitution as stated below:

We solve for $x_3$ from Equation (4.11(c)), then solve for $x_2$ from Equation (4.11(b)) and finally solve for $x_1$ from Equation (4.11(a)). This systematic Gaussian elimination procedure can be written in matrix notation in a compact form, as shown below.

NOTES

(*i*) We write the augmented matrix, $[A : b]$ and the multipliers on the left.

$$\begin{matrix} \\ m_2 = -a_{21}/a_{11} \\ m_3 = -a_{31}/a_{11} \end{matrix} \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{bmatrix}. \xrightarrow[R_2 - m_2R_1 \text{ and } R_3 - m_3R_1]{\text{Perform row operators}}$$

(*ii*) Then we write the transformed 2nd and 3rd rows after the elimination of $x_1$ by row operations $[(m_2 \times 1\text{st row} + 2\text{nd row})$ and $(m_3 \times 1\text{st row} + 3\text{rd row})]$ as new 2nd and 3rd rows along with the multiplier on the left.

$$m_4 = -a_{31}^{(1)}/a_{22}^{(1)} \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ & a_{22}^{(1)} & a_{23}^{(1)} & b_2^{(1)} \\ & a_{32}^{(1)} & a_{33}^{(1)} & b_3^{(1)} \end{bmatrix}. \xrightarrow[R_3 - m_4R_2]{\text{Perform}}$$

(*iii*) Finally, we get the upper triangular transformed augmented matrix as given below.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ & a_{22}^{(1)} & a_{23}^{(1)} & b_2^{(1)} \\ & & a_{33}^{(2)} & b_3 \end{bmatrix} \qquad (4.12)$$

***Notes:***

1. The above procedure can be easily extended to a system of $n$ unknowns, in which case, we have to perform a total of $(n–1)$ steps for the systematic elimination to get the final upper triangular matrix.
2. The condition to be satisfied for using this elimination is that the first diagonal elements at each step must not be zero. These diagonal elements $[a_{11}, a_{22}^{(1)}, a_{33}^{(2)}$, etc.] are called pivot. If the pivot is zero at any stage, the method fails. However, we can rearrange the rows so that none of the pivots is zero, at any stage.

**Example 3:** Solve the following system by Gauss elimination method:

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 0 \\ 2x_1 + 2x_2 + 3x_3 &= 3 \\ -x_1 - 3x_2 &= 2 \end{aligned}$$

Show the computations by augmented matrix representation.

**Solution:** The augmented matrix of the system is,

$$\begin{bmatrix} 1 & 2 & 1 & : & 0 \\ 2 & 2 & 3 & : & 3 \\ -1 & -3 & 0 & : & 2 \end{bmatrix}$$

**NOTES**

**Step 1:** For elimination of $x_1$ from the 2nd and 3rd equations we multiply the first equation by –2 and 1 successively and add them to the 2nd and 3rd equation. The result is shown in the augmented matrix below.

$$\begin{matrix} \\ -2 \\ 1 \end{matrix}\begin{bmatrix} 1 & 2 & 1 & : & 0 \\ 0 & -2 & 1 & : & 3 \\ 0 & -1 & 1 & : & 2 \end{bmatrix}$$

**Step 2:** For elimination of $x_2$ from the third equation we multiply the second equation by $-\frac{1}{2}$ and add it to the third equation. The result is shown in the augmented matrix below.

$$\begin{matrix} \\ -1/2 \\ \\ \end{matrix}\begin{bmatrix} 1 & 2 & 1 & : & 0 \\ 0 & -2 & 1 & : & 3 \\ 0 & 0 & 1/2 & : & 1/2 \end{bmatrix}$$

**Step 3:** The upper triangular system is now solved by back-substitution, giving

$$x_1 = 1, x_2 = -1, x_3 = 1$$

**Gauss-Jordan Elimination Method:** The Gauss-Jordan elimination method is a variation of the Gaussian elimination method. In this method, the augmented coefficient matrix is transformed by row operations such that the coefficient matrix reduces to the identity matrix. The solution of the system is then directly obtained as the reduced augmented column of the transformed augmented matrix. We explain the method with a system of three equations given by,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad (4.13)$$

The augmented matrix is,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & : & b_1 \\ a_{21} & a_{22} & a_{23} & : & b_2 \\ a_{31} & a_{32} & a_{33} & : & b_3 \end{bmatrix} \quad (4.14)$$

We assume that $a_{11}$ is non-zero. If, however, $a_{11}$ is zero, we can interchange rows so that $a_{11}$ is non-zero in the resulting system.

The first step is to divide the first row by $a_{11}$ and then eliminating $x_1$ from 2nd and 3rd equations by row operations of multiplying the reduced first row by $a_{21}$ and subtracting from the second row and next multiplying the reduced first row by $a_{31}$ and subtracting from the third row. This is shown in matrix transformations given below.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & : & b_1 \\ a_{21} & a_{22} & a_{23} & : & b_2 \\ a_{31} & a_{32} & a_{33} & : & b_3 \end{bmatrix} \xrightarrow[R_1/a_{11}]{} \begin{bmatrix} 1 & a'_{12} & a'_{13} & : & b'_1 \\ a_{21} & a_{22} & a'_{23} & : & b_2 \\ a_{31} & a_{32} & a'_{33} & : & b_3 \end{bmatrix} \xrightarrow[R_3-R_1\ a_{31}]{R_2-R_1\ a_{21}} \begin{bmatrix} 1 & a'_{12} & a'_{13} & : & b'_1 \\ 0 & a'_{22} & a'_{23} & : & b'_2 \\ 0 & a'_{32} & a'_{33} & : & b'_3 \end{bmatrix}$$

**NOTES**

where,

$$a'_{12} = a_{12}/a_{11},\ a'_{13} = a_{13}/a_{11},\ b'_1 = b_1/a_{11}$$
$$a'_{22} = a_{22} - a_{21}\ a'_{12},\ a'_{23} = a_{23} - a_{21}\ a'_{13},\ b'_2 = b_2 - a_{21}b'_1$$
$$a'_{32} = a_{32} - a_{31}\ a'_{12},\ a'_{33} = a_{33} - a_{31}\ a'_{13},\ b'_3 = b_3 - a_{31}b'_1$$

Now considering $a'_{22}$ as the non-zero pivot, we first divide the second row by $a'_{22}$ and then multiply the reduced second row by $a'_{12}$ and subtract it from the first row and also multiply the reduced second row by $a'_{32}$ and subtracting it from the third row. The operations are shown below in matrix notation.

$$\begin{bmatrix} 1 & a'_{12} & a'_{13} & : & b'_1 \\ 0 & a'_{22} & a'_{23} & : & b'_2 \\ 0 & a'_{32} & a'_{33} & : & b'_3 \end{bmatrix} \xrightarrow{R_2/a'_{22}} \begin{bmatrix} 1 & a'_{12} & a'_{13} & : & b'_1 \\ 0 & 1 & a''_{23} & : & b''_2 \\ 0 & a'_{32} & a'_{33} & : & b'_3 \end{bmatrix} \xrightarrow{R_1-R_2\ a'_{12} \text{ and } R_3-R_2\ a'_{32}} \begin{bmatrix} 1 & 0 & a''_{13} & : & b''_1 \\ 0 & 1 & a''_{23} & : & b''_2 \\ 0 & 0 & a''_{33} & : & b''_3 \end{bmatrix}$$

where

$$a''_{13} = a'_{13} - a'_{12}\ a''_{23},\ b''_{11} = b'_1 - a'_{12}b''_2$$
$$a''_{23} = a'_{23}/a'_{22},\ b''_2 = b'_2/a'_{22}$$
$$a''_{33} = a'_{33} - a''_{23}\ a'_{32},\ b''_3 = b'_3 - a'_{32}b''_2$$

Finally, the third row elements are divided by $a''_{33}$ and then the reduced third row is multiplied by $a''_{13}$ and subtracted from the first row and also the reduced third row is multiplied by $a''_{23}$ and subtracted from the second row. This is again shown in matrix notation below.

$$\begin{bmatrix} 1 & 0 & a''_{13} & : & b''_1 \\ 0 & 1 & a''_{23} & : & b''_2 \\ 0 & 0 & a''_{33} & : & b''_3 \end{bmatrix} \xrightarrow[R_3/a''_{33}]{} \begin{bmatrix} 1 & 0 & a''_{13} & : & b''_1 \\ 0 & 1 & a''_{23} & : & b''_2 \\ 0 & 0 & 1 & : & b'''_3 \end{bmatrix} \xrightarrow[R_2 - a''_{23}R_3]{R_1 - a''_{25}R_3} \begin{bmatrix} 1 & 0 & 0 & : & b''_1 \\ 0 & 1 & 0 & : & b''_2 \\ 0 & 0 & 1 & : & b''_3 \end{bmatrix}$$

where $b''_1 = b''_1 - a''_{13}b'''_3$ , $b'''_2 = b''_2 - a''_{23}b'''_3$, $b'''_3 = b''_3 / a''_{33}$

Finally, the solution of the system is given by the reduce augmented column, i.e., $x_1 = b'''_1,\ x_2 = b'''_2$ and $x_3 = b'''_3.$

We illustrate the elimination procedure with an example using augmented matrix,

$$\begin{bmatrix} 2 & 2 & 4 & : & 18 \\ 1 & 3 & 2 & : & 13 \\ 3 & 1 & 3 & : & 14 \end{bmatrix}$$

First, we divide the first row by 2 then subtract the reduced first row from 2nd row and also multiply the first row by 2 and then subtract from the third. The results are shown below:

$$\begin{bmatrix} 1 & 2 & 4 & 18 \\ 1 & 3 & 2 & 13 \\ 3 & 1 & 3 & 14 \end{bmatrix} \xrightarrow{R_1/2} \begin{bmatrix} 1 & 1 & 2 & : & 9 \\ 1 & 3 & 2 & : & 13 \\ 3 & 1 & 3 & : & 14 \end{bmatrix} \xrightarrow[R_3+2R_2]{R_2-R_1} \begin{bmatrix} 1 & 1 & 2 & : & 9 \\ 0 & 2 & 0 & : & 4 \\ 0 & -2 & -3 & : & -13 \end{bmatrix}$$

Next considering 2nd row, we reduce the second column to [0, 1, 0] by row operations shown below:

$$\begin{bmatrix} 1 & 1 & 2 & : & 9 \\ 0 & 2 & 0 & : & 4 \\ 0 & -2 & -3 & : & -13 \end{bmatrix} \xrightarrow{R_2/2} \begin{bmatrix} 1 & 1 & 2 & : & 9 \\ 0 & 1 & 0 & : & 2 \\ 0 & -2 & -3 & : & -13 \end{bmatrix} \xrightarrow[R_3+2R_2]{R_1-R_2} \begin{bmatrix} 1 & 0 & 2 & : & 7 \\ 0 & 1 & 0 & : & 2 \\ 0 & 0 & -3 & : & -9 \end{bmatrix}$$

Finally, dividing the third row by –3 and then subtracting from the first row the elements of the third row multiplied by 2, the result is shown below:

$$\begin{bmatrix} 1 & 0 & 2 & : & 7 \\ 0 & 1 & 0 & : & 2 \\ 0 & 0 & -3 & : & -9 \end{bmatrix} \xrightarrow{R_3/(-3)} \begin{bmatrix} 1 & 0 & 2 & : & 7 \\ 0 & 1 & 0 & : & 2 \\ 0 & 0 & 1 & : & 3 \end{bmatrix} \xrightarrow{R_1-2R_3} \begin{bmatrix} 1 & 0 & 0 & : & 1 \\ 0 & 1 & 0 & : & 2 \\ 0 & 0 & 1 & : & 3 \end{bmatrix}$$

Hence the solution of the system is $x_1 = 1, x_2 = 2, x_3 = 3$.

**Example 4:** Solve the following system by Gauss-Jordan elimination method:

$$\begin{bmatrix} 3 & 18 & 9 \\ 2 & 3 & 3 \\ 4 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 18 \\ 117 \\ 283 \end{bmatrix}$$

**Solution:** We consider the augmented matrix and solve the system by Gauss-Jordan elimination method. The computations are shown in compact matrix notation as given below. The augmented matrix is,

$$\begin{bmatrix} 3 & 18 & 9 & : & 18 \\ 2 & 3 & 3 & : & 117 \\ 4 & 1 & 2 & : & 283 \end{bmatrix}$$

**Step 1:** The pivot is 3 in the first column. The first column is transformed into $[1, 0, 0]^T$ by row operations shown below:

$$\begin{bmatrix} 3 & 18 & 9 & : & 18 \\ 2 & 3 & 3 & : & 117 \\ 4 & 1 & 2 & : & 283 \end{bmatrix} \xrightarrow{R_1/3} \begin{bmatrix} 1 & 6 & 3 & 6 \\ 2 & 3 & 3 & 117 \\ 4 & 1 & 2 & 283 \end{bmatrix} \xrightarrow[R_3-4R_1]{R_2-2R_1} \begin{bmatrix} 1 & 6 & 3 & : & 6 \\ 0 & -9 & -3 & : & 105 \\ 0 & -23 & -10 & : & 259 \end{bmatrix}$$

**Step 2:** The second column is transformed into [0, 1, 0] by row operations shown below:

$$\begin{bmatrix} 1 & 6 & 3 & : & 6 \\ 0 & -9 & -3 & : & 105 \\ 0 & -23 & -10 & : & 259 \end{bmatrix} \xrightarrow{-R_2/9} \begin{bmatrix} 1 & 6 & 3 & : & 6 \\ 0 & 1 & 1/3 & : & -35/3 \\ 0 & -23 & -10 & : & 259 \end{bmatrix} \xrightarrow[R_3+23R_2]{R_1-6R_2} \begin{bmatrix} 1 & 0 & 1 & : & 76 \\ 0 & 1 & 1/3 & : & -35/3 \\ 0 & 0 & -7/3 & : & 28/3 \end{bmatrix}$$

**Step 3:** The third column is transformed into $[0\ 0\ 1]^T$ by row operations shown below:

$$\begin{bmatrix} 1 & 0 & 1 & : & 76 \\ 0 & 1 & 1/3 & : & -35/3 \\ 0 & 0 & -7/3 & : & 28/3 \end{bmatrix} \xrightarrow{R_3/(-7/3)} \begin{bmatrix} 1 & 0 & 1 & : & 76 \\ 0 & 1 & 1/3 & : & -35/3 \\ 0 & 0 & 1 & : & 4 \end{bmatrix} \xrightarrow[R_2-R_3/3]{R_1-R_3} \begin{bmatrix} 1 & 0 & 0 & : & 72 \\ 0 & 1 & 0 & : & -13 \\ 0 & 0 & 1 & : & 4 \end{bmatrix}$$

Hence the solution of the system is $x_1 = 72, x_2 = -13, x_3 = 4$.

### 4.2.3 Iterative Methods

We can use iteration methods to solve a system of linear equations when the coefficient matrix is diagonally dominant. This is ensured by the set of sufficient conditions given as follows,

$$\sum_{j=1,\ j\neq i}^{n} |a_{ij}| < |a_{ii}|, \text{ for } i = 1, 2, ..., n \tag{4.15}$$

An alternative set of sufficient conditions is,

$$\sum_{i=1,\ i\neq j}^{n} |a_{ij}| < |a_{jj}|, \text{ for } j = 1, 2, ..., n \tag{4.16}$$

There are two forms of iteration methods termed as Jacobi iteration method and Gauss-Seidel iteration method.

**Jacobi Iteration Method:** Consider a system of $n$ linear equations,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ........ + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + ........ + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + ........ + a_{3n}x_n &= b_3 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + ........ + a_{nn}x_n &= b_n \end{aligned}$$

The diagonal elements $a_{ii}$, $i = 1, 2, ..., n$ are non-zero and satisfy the set of sufficient conditions stated earlier. When the system of equations do not satisfy these conditions, we may rearrange the system in such a way that the conditions hold.

In order to apply the iteration we rewrite the equations in the following form:

NOTES

$$
\begin{aligned}
x_1 &= (b_1 - a_{12}x_2 - a_{13}x_3 - ... - a_{1n}\, x_n)/a_{11}\\
x_2 &= (b_2 - a_{21}x_1 - a_{23}x_3 - ... - a_{2n}\, x_n)/a_{22}\\
x_3 &= (b_3 - a_{31}x_1 - a_{32}x_2 - ... - a_{3n}\, x_n)/a_{33}\\
&\text{.......................................................}\\
x_n &= (b_n - a_{n1}x_1 - a_{n2}x_2 - ... - a_{nn-1}\, x_{n-1})/a_{nn}
\end{aligned}
$$

To start the iteration we make an initial guess of the unknowns as $x_1^{(0)}, x_2^{(0)}, x_3^{(0)},..., x_n^{(0)}$ (initial guess may be taken to be zero).

Successive approximations are computed using the equations,

$$
\begin{aligned}
x_1^{(k+1)} &= (b_1 - a_{12}\, x_2^{(k)} - a_{13}\, x_3^{(k)} - ... - a_{1n}\, x_n^{(k)})/a_{11}\\
x_2^{(k+1)} &= (b_2 - a_{21}\, x_1^{(k)} - a_{23}\, x_3^{(k)} - ... - a_{2n}\, x_n^{(k)})/a_{22}\\
x_3^{(k+1)} &= (b_3 - a_{31}\, x_1^{(k)} - a_{32}\, x_2^{(k)} - ... - a_{3n}\, x_n^{(k)})/a_{33}\\
&\text{......................................................................}\\
x_n^{(k+1)} &= (b_n - a_{n1}\, x_1^{(k)} - a_{n2}\, x_2^{(k)} - ... - a_{nn-1}\, x_{n-1}^{(k)})/a_{nn}
\end{aligned}
\tag{4.17}
$$

where $k = 0, 1, 2, ...$

The iterations are continued till the desired accuracy is achieved. This is checked by the relations,

$$\left| x_i^{(k+1)} - x_i^{(k)} \right| < \varepsilon, \text{ for } i = 1, 2, ..., n \tag{4.18}$$

**Jacobi Iterative Algorithm**

Choose an initial guess $x^{(0)}$ to the solution $x$.

**for** $k = 1, 2, \ldots$

**for** $i = 1, 2, \ldots, n$

$\bar{x}_i = 0$

**for** $j = 1, 2, \ldots, i-1, i+1, \ldots, n$

$\bar{x}_i = \bar{x}_i + a_i, j\, x_j^{(k-1)}$

**end**

$\bar{x}_i = (b_i - \bar{x}_i)/a_{i,i}$

**end**

$x^{(k)} = \bar{x}$

check convergence; continue if necessary

**end**

**Gauss-Seidel Iteration Method:** This is a simple modification of the Jacobi iteration. In this method, at any stage of iteration of the system, the improved values of the unknowns are used for computing the components of the unknown vector. The iteration equations given below are used in this method,

$$\begin{aligned}
x_1^{(k+1)} &= (b_1 - a_{12}\, x_2^{(k)} - a_{13}x_3^{(k)} - \ldots - a_{1n}\ x_n^{(k)})/a_{11} \\
x_2^{(k+1)} &= (b_2 - a_{21}\, x_1^{(k+1)} - a_{23}x_3^{(k)} - \ldots - a_{2n}\ x_n^{(k)})/a_{22} \\
x_3^{(k+1)} &= (b_3 - a_{31}\, x_1^{(k+1)} - a_{32}x_2^{(k+1)} - \ldots - a_{3n}\ x_n^{(k)})/a_{33} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_n^{(k+1)} &= (b_n - a_{n1}\, x_1^{(k+1)} - a_{n2}^{(k+1)} x_2(k+1) - \ldots - a_{nn-1}\ x_{n-1}^{(k+1)})/a_{nn}
\end{aligned} \qquad (4.19)$$

It is clear from above that for computing $x_2^{(k+1)}$, the improved value of $x_1^{(k+1)}$ is used instead of $x_1^{(k)}$; and for computing $x_3^{(k+1)}$, the improved values $x_1^{(k+1)}$ and $x_2^{(k+1)}$ are used. Finally, for computing $x_n^{(k)}$, improved values of all the components $x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}$ are used. Further, as in the Jacobi iteration, the iterations are continued till the desired accuracy is achieved.

**Example 5:** Solve the following system by Gauss-Seidel iterative method correct upto four significant digits.

$$\begin{aligned}
10x_1 - 2x_2 - x_3 - x_4 &= 3 \\
-2x_1 + 10x_2 - x_3 - x_4 &= 15 \\
-x_1 - x_2 + 10x_3 - 2x_4 &= 27 \\
-x_1 - x_2 - 2x_3 + 10x_4 &= -9
\end{aligned}$$

**Solution:** The given system is clearly having diagonally dominant coefficient matrix,

i.e., $$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, 2, \ldots, n$$

Hence, we can employ Gauss-Seidel iteration method, for which we rewrite the system as,

$$\begin{aligned}
x_1^{(k+1)} &= 0.3 + 0.2\, x_2^{(k)} + 0.1\ x_3^{(k)} + 0.1\ x_4^{(k)} \\
x_2^{(k+1)} &= 1.5 + 0.2\, x_1^{(k+1)} + 0.1\ x_3^{(k)} + 0.1\ x_4^{(k)} \\
x_3^{(k+1)} &= 2.7 + 0.1 x_1^{(k+1)} + 0.1\ x_2^{(k+1)} + 0.2\ x_4^{(k)} \\
x_4^{(k+1)} &= -0.9 + 0.1 x_1^{(k+1)} + 0.1\ x_2^{(k+1)} + 0.2\ x_3^{(k+1)}
\end{aligned}$$

We start the iteration with,

$$x_1^{(0)} = 0.3,\ x_2^{(0)} = 1.5,\ x_3^{(0)} = 2.7,\ x_4^{(0)} = -0.9$$

The results of successive iterations are given in the table below.

| $k$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| 1 | 0.72 | 1.824 | 2.774 | –0.0196 |
| 2 | 0.9403 | 1.9635 | 2.9864 | –0.0125 |
| 3 | 0.09901 | 1.9954 | 2.9960 | –0.0023 |
| 4 | 0.9984 | 1.9990 | 2.9993 | –0.0004 |
| 5 | 0.9997 | 1.9998 | 2.9998 | –0.0003 |
| 6 | 0.9998 | 1.9998 | 2.9998 | –0.0003 |
| 7 | 1.0000 | 2.0000 | 3.0000 | 0.0000 |

Hence the solution correct to four significant figures is $x_1 = 1.0000$, $x_2 = 2.000$, $x_3 = 3.000$, $x_4 = 0.000$.

**Example 6:** Solve the following system by Gauss-Seidel iteration method.

$$20x_1 + 2x_2 + x_3 = 30$$
$$x_1 - 40x_2 + 3x_3 = -75$$
$$2x_1 - x_2 + 10x_2 = 30$$

Give the solution correct upto three significant figures.

**Solution:** It is evident that the coefficient matrix is diagonally dominant and the sufficient conditions for convergence of the Gauss-Seidel iterations are satisfied, since

$$|a_{11}| = 20 \geq |a_{12}| + |a_{13}| = 3$$
$$|a_{22}| = 40 \geq |a_{21}| + |a_{23}| = 4$$
$$|a_{33}| = 10 \geq |a_{31}| + |a_{32}| = 3$$

For starting the iterations, we rewrite the equations as,

$$x_1 = \frac{1}{20}(30 - 2x_2 - x_3)$$
$$x_2 = \frac{1}{40}(75 + x_1 + 3x_3)$$
$$x_3 = \frac{1}{10}(30 - 2x_1 + x_2)$$

The initial approximate solution is taken as,

$$x_1^{(0)} = 1.5, \quad x_2^{(0)} = 2.0, \quad x_3^{(0)} = 3.0$$

The first iteration gives,

$$x_1^{(1)} = \frac{1}{20}(30 - 2\times 2.0 - 3.0) = 1.15$$
$$x_2^{(1)} = \frac{1}{40}(75 + 1.15 + 3\times 3.0) = 2.14$$
$$x_3^{(1)} = \frac{1}{10}(30 - 2\times 1.15 + 2.14) = 2.98$$

NOTES

The second iteration gives,

$$x_1^{(2)} = \frac{1}{20}(30 - 2\times 2.14 - 2.98) = 1.137$$

$$x_2^{(2)} = \frac{1}{40}(75 + 1.137 + 3\times 2.98) = 2.127$$

$$x_3^{(2)} = \frac{1}{10}(30 - 2\times 1.137 + 2.127) = 2.986$$

The third iteration gives,

$$x_1^{(3)} = \frac{1}{20}(30 - 2\times 2.127 - 2.986) = 1.138$$

$$x_2^{(3)} = \frac{1}{40}(75 + 1.138 + 3\times 2.986) = 2.127$$

$$x_3^{(3)} = \frac{1}{10}(30 - 2\times 1.138 + 2.127) = 2.985$$

Thus the solution correct to three significant digits can be written as $x_1 = 1.14$, $x_2 = 2.13$, $x_3 = 2.98$.

**Example 7:** Solve the following system correct to three significant digits, using Jacobi iteration method.

$$10x_1 + 8x_2 - 3x_3 + x_4 = 16$$
$$3x_1 - 4x_2 + 10x_3 + x_4 = 10$$
$$2x_1 + 10x_2 + x_3 - 4x_4 = 9$$
$$2x_1 + 2x_2 - 3x_3 + 10x_4 = 11$$

**Solution:** The system is first rearranged so that the coefficient matrix is diagonally dominant. The equations are rewritten for starting Jacobi iteration as,

$$x_1^{(k+1)} = 1.6 - 0.8\ x_2^{(k)} + 0.3\ x_3^{(k)} - 0.1\ x_4^{(k)}$$
$$x_2^{(k+1)} = 0.9 - 0.2\ x_1^{(k)} + 0.1\ x_3^{(k)} - 0.4\ x_4^{(k)}$$
$$x_3^{(k+1)} = 1.0 - 0.3\ x_1^{(k)} + 0.4\ x_2^{(k)} - 0.1\ x_4^{(k)}$$
$$x_4^{(k+1)} = 1.1 - 0.2\ x_1^{(k)} + 0.2\ x_2^{(k)} - 0.3\ x_3^{(k)}, \quad \text{where} \quad k = 0,\ 1,\ 2, ...$$

The initial guess of solution is taken as,

$$x_1^{(0)} = 1.6, \quad x_2^{(0)} = 0.9, \quad x_3^{(0)} = 1.0, \quad x_4^{(0)} = 1.1$$

The results of successive iterations computed by Jacobi iterations are given in the following table:

| $k$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| 1 | 1.07 | 0.92 | 0.77 | 0.90 |
| 2 | 1.050 | 0.969 | 0.957 | 0.933 |
| 3 | 1.0186 | 0.9765 | 0.9928 | 0.9923 |
| 4 | 1.0174 | 0.9939 | 0.9858 | 0.9989 |
| 5 | 0.9997 | 0.9975 | 0.9925 | 0.9974 |
| 6 | 1.0001 | 0.9997 | 0.9994 | 0.9984 |
| 7 | 1.0002 | 0.9998 | 1.0001 | 0.9999 |

Thus the solution correct to three significant digits is $x_1 = 1.000$, $x_2 = 1.000$, $x_4 = 1.000$.

**Algorithm:** Solution of a system of equations by Gauss-Seidel iteration method.

**Step 1:** Input elements $a_{ij}$ of augmented matrix for $i = 1$ to $n$, next, $j = 1$ to $n + 1$.

**Step 2:** Input epsilon, maxit [epsilon is desired accuracy, maxit is maximum number of iterations]

**Step 3:** Set $x_i = 0$, for $i = 1$ to $n$

**Step 4:** Set big = 0, sum = 0, $j = 1$, $k = 1$, iter = 0

**Step 5:** Check if $k \neq j$, set sum = sum + $a_{jk} x_k$

**Step 6:** Check if $k < n$, set $k = k + 1$, go to Step 5 else go to next step

**Step 7:** Compute temp = $(a_{jn+1} - \text{sum}) / a_{jj}$

**Step 8:** Compute relerr = abs $(x_j - \text{temp})$ / temp

**Step 9:** Check if big < relerr then big = relerr

**Step 10:** Set $x_j$ = temp

**Step 11:** Set $j = j + 1$, $k = 1$

**Step 12:** Check if $j \leq n$ to Step 5 else go to next step

**Step 13:** Check if relerr < epsilon then {write iterations converge, and write $x_j$ for $j = 1$ to $n$ go to Step 15} else if iter < maxit iter = iter + 1 go to Step 5

**Step 14:** Write 'iterations do not converge in', maxit 'iteration'

**Step 15:** Write $x_j$ for $j = 1$ to $n$

**Step 16:** End

### 4.2.4 Computation of the Inverse of a Matrix by using Gaussian Elimination Method

The inverse matrix $B$ of a given square matrix $A$ satisfy the relation,

$$A \,.\, B = I$$

where $I$ is the unit matrix of the same order as that of $A$. In order to determine the elements $b_{ij}$ of the matrix $B$, we can employ row operations as in Gaussian elimination. We explain the method for a 2 × 3 matrix as given below. We can write the above relation in detail as,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

By using the definition of matrix multiplication we can write that the above relation equivalent to the following three systems of linear equations.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Thus by solving each of the above systems we shall get the three columns of the inverse matrix $B = A^{-1}$. Since, the coefficient matrix is the same for each of the three systems, we can apply Gauss elimination to all the three systems simultaneously. We consider for this the following augmented matrix:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & : & 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & : & 0 & 1 & 0 \\ a_{31} & a_{32} & a_{33} & : & 0 & 0 & 1 \end{bmatrix}$$

We employ Gauss elimination to this augmented matrix. At the end of 1st stage we get,

$$\xrightarrow[R_3 - (a_{31}/a_{11})R_1]{R_2 - (a_{21}/a_{11})R_1} \begin{bmatrix} a_{11} & a_{12} & a_{13} & : & 1 & 0 & 0 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & : & -a_{21}/a_{11} & 1 & 0 \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & : & -a_{31}/a_{11} & 0 & 1 \end{bmatrix}$$

where

$$a_{22}^{(1)} = a_{22} - (a_{21}/a_{11})a_{12},\ a_{23}^{(1)} = a_{23} - (a_{21}/a_{11})a_{13}$$

$$a_{32}^{(1)} = a_{32} - (a_{31} - a_{11})a_{12},\ a_{33}^{(1)} = a_{33} - (a_{31}/a_{11})a_{13}$$

Similarly, at the end of the second stage, we have

$$R_3 - \left(a_{23}^{(1)}/a_{22}^{(1)}\right) a_{32}^{(1)} \longrightarrow \begin{bmatrix} a_{11} & a_{12} & a_{13} & : & 1 & 0 & 0 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & : & c_{21} & 1 & 0 \\ 0 & 0 & a_{33}^{(2)} & : & c_{31} & c_{32} & 1 \end{bmatrix}$$

where

$$a_{33}^{(2)} = a_{33}^{(1)} - \left(a_{32}^{(1)}/a_{22}^{(1)}\right) \ a_{23}^{(1)},\ \ c_{21} = -(a_{21}/a_{12})$$

$$= \left(a_{21} \quad a_{32}^{(1)}/a_{11} \quad a_{22}^{(1)}\right) - (a_{31}/a_{11}),\ \ c_{32} = -\left(a_{32}^{(1)}/a_{22}^{(1)}\right)$$

By back-substitution process, we get the elements of the inverse matrix, by solving the three systems corresponding to the three columns of the reduced augmented part, i.e.,

$$\begin{bmatrix} 1 & 0 & 0 \\ c_{11} & 1 & 0 \\ c_{31} & c_{32} & 1 \end{bmatrix}$$

We illustrate the method by an example given below.

**Example 8:** Find the inverse of the following matrix $A$ by Gaussian elimination method.

$$A = \begin{bmatrix} 2 & 3 & -1 \\ 4 & 4 & -3 \\ 2 & -3 & 1 \end{bmatrix}$$

**Solution:** We consider the following augmented matrix:

$$[A.I] = \begin{bmatrix} 2 & 3 & -1 & : & 1 & 0 & 0 \\ 4 & 4 & -3 & : & 0 & 1 & 0 \\ 2 & -3 & 1 & : & 0 & 0 & 1 \end{bmatrix}$$

**NOTES**

Using Gaussian elimination to this augmented matrix, we get the following at the end of first step:

$$\xrightarrow[R_3 - R_1]{R_2 - 2R_1} \begin{bmatrix} 2 & 3 & -1 & : & 1 & 0 & 0 \\ 0 & -2 & -1 & : & -2 & 1 & 0 \\ 0 & -6 & 2 & : & -1 & 0 & 1 \end{bmatrix}$$

Similarly, at the end of 2nd step we get,

$$\xrightarrow[R_3 - 3R_2]{} \begin{bmatrix} 2 & 3 & -1 & : & 1 & 0 & 0 \\ 0 & -2 & -1 & : & -2 & 1 & 0 \\ 0 & 0 & 5 & : & 5 & -3 & 1 \end{bmatrix}$$

Thus, we get the three columns of inverse matrix by solving the following three systems:

$$\begin{bmatrix} 2 & 3 & -1 & : & 1 \\ 0 & -2 & -1 & : & -2 \\ 0 & 0 & 5 & : & 5 \end{bmatrix} \begin{bmatrix} 2 & 3 & -1 & : & 0 \\ 0 & -2 & -1 & : & 1 \\ 0 & 0 & 5 & : & -3 \end{bmatrix} \begin{bmatrix} 2 & 3 & -1 & : & 0 \\ 0 & -2 & -1 & : & 0 \\ 0 & 0 & 5 & : & 1 \end{bmatrix}$$

The solution of the three are easily derived by back-substitution, which give the three columns of the inverse matrix given below:

$$\begin{bmatrix} 1/4 & 0 & 1/4 \\ 1/2 & -1/5 & -1/10 \\ 1 & -3/5 & 1/5 \end{bmatrix}$$

We can also employ Gauss-Jordan elimination to compute the inverse matrix. This is illustrated by the following example:

**Example 9:** Compute the inverse of the following matrix by Gauss-Jordan elimination.

$$A = \begin{bmatrix} 2 & 3 & -1 \\ 4 & 4 & -3 \\ 2 & -3 & 1 \end{bmatrix}$$

**Solution:** We consider the augmented matrix $[A : I]$,

$$[A:I] = \begin{bmatrix} 2 & 3 & -1 & : & 1 & 0 & 0 \\ 4 & 4 & -3 & : & 0 & 1 & 0 \\ 2 & -3 & 1 & : & 0 & 0 & 1 \end{bmatrix} \xrightarrow[R_1/2]{} \begin{bmatrix} 1 & 3/2 & -1/2 & : & 1/2 & 0 & 0 \\ 4 & 4 & -3 & : & 0 & 1 & 0 \\ 2 & -3 & 1 & : & 0 & 0 & 1 \end{bmatrix}$$

$$\xrightarrow[R_2 - 4R_1]{R_3 - 2R_1} \begin{bmatrix} 1 & 3/2 & -1/2 & : & 1/2 & 0 & 0 \\ 0 & -2 & -1 & : & -2 & 1 & 0 \\ 0 & -6 & 2 & : & -1 & 0 & 1 \end{bmatrix} \xrightarrow[R_2/-2]{} \begin{bmatrix} 1 & 3/2 & -1/2 & : & 1/2 & 0 & 0 \\ 0 & 1 & +1/2 & : & 1 & -1/2 & 0 \\ 0 & -6 & 2 & : & -1 & 0 & 1 \end{bmatrix}$$

$$\xrightarrow[R_3 + 6R_2]{R_1 - 3R_2/2} \begin{bmatrix} 1 & 0 & -5/4 & : & -1 & 3/4 & 0 \\ 0 & 1 & 1/2 & : & 1 & -1/2 & 0 \\ 0 & 0 & 5 & : & 5 & -3 & 1 \end{bmatrix} \xrightarrow[R_3/5]{} \begin{bmatrix} 1 & 0 & -5/4 & : & -1 & 3/4 & 0 \\ 0 & 1 & 1/2 & : & 1 & -1/2 & 0 \\ 0 & 0 & 1 & : & 1 & -3/5 & 1/5 \end{bmatrix}$$

$$\xrightarrow[R_2 - 1R_3/2]{R_1 + 5R_3/4} \begin{bmatrix} 1 & 0 & 0 & : & 1/4 & 0 & 1/4 \\ 0 & 1 & 0 & : & 1/2 & -1/5 & -1/10 \\ 0 & 0 & 1 & : & 1 & -3/5 & 1/5 \end{bmatrix}$$

which gives $A^{-1} = \begin{bmatrix} 1/4 & 0 & 1/4 \\ 1/2 & -1/5 & -1/10 \\ 1 & -3/5 & 1/5 \end{bmatrix}$

**NOTES**

**Check Your Progress**

1. When is the system of equation homogenous and when non-homogenous?
2. Explain Gauss elimination method.
3. Explain Gauss-Jordan elimination method.
4. Why are iterative methods used?
5. Explain Gauss-Seidel iteration method.

## 4.3 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The system of equations $Ax = b$ is termed as a homogeneous one if all the elements in the column vector $b$ are zero. Otherwise, the system is termed as a non-homogeneous one.
2. The gauss elemination method consists in systematic elimination of the unknowns so as to reduce the coefficient matrix into an upper triangular system, which is then solved by the procedure of back-substitution.
3. The Gauss-Jordan elimination method is a variation of the Gaussian elimination method. In this method, the augmented coefficient matrix is transformed by row operations such that the coefficient matrix reduces to the Identity matrix. The solution of the system is then directly obtained as the reduced augmented column of the transformed augmented matrix.
4. We can use iteration methods to solve a system of linear equations when the coefficient matrix is diagonally dominant.
5. The Gauss-Seidel iteration is a simple modification of the Jacobi iteration. In this method, at any stage of iteration of the system, the improved values of the unknowns are used for computing the components of the unknown vector.

## 4.4 SUMMARY

- Many engineering and scientific problems require the solution of a system of linear equations.
- The system of equations is termed as a homogeneous one if all the elements in the column vector $b$ of the equation $Ax = b$, are zero.
- Cramer's rule and matrix inversion method are two classical methods to solve the system of equations.
- If $D = |A|$ be the determinant of the coefficient matrix $A$ and $D_i$ is the determinant obtained by replacing the $i$th column of $D$ by the column vector $b$, then the Cramer's rule gives the solution vector $x$ by the equations,

$$x_i = \frac{D_i}{D} \text{ for } i = 1, 2, \ldots, n.$$

NOTES

- Gaussian elimination method consists in systematic elimination of the unknowns so as to reduce the coefficient matrix into an upper triangular system, which is then solved by the procedure of back-substitution.
- In Gauss-Jordan elimination, the augmented matrix is transformed by row operations such that the coefficient matrix reduces to the identity matrix.
- We can use iteration methods to solve a system of linear equations when the coefficient matrix is diagonally dominant.
- There are two forms of iteration methods termed as Jacobi iteration method and Gauss-Seidel iteration method.
- Gaussian elimination can be used to compute the inverse of a matrix.

## 4.5 KEY WORDS

- **Homogenous equation:** In this system of equations, all the elements in the column vector $b$ of the equation $Ax = b$, are zero.
- **Gaussian elimination:** It is the systematic elimination of the unknowns so as to reduce the coefficient matrix into an upper triangular system, which is then solved by the procedure of back-substitution.
- **Gauss-Seidel iteration:** In this method, at any stage of iteration of the system, the improved values of the unknowns are used for computing the components of the unknown vector.

## 4.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

**Short-Answer Questions**

1. Define the system of linear equations.
2. How many determinants $d_0$ we have to compute in Cramer's rule?
3. What is the basic difference between Gaussian elimination and Gauss-Jordan elimination method?
4. What are iterative methods?
5. State an application of Gaussian elimination method.

**Long-Answer Questions**

1. Use Cramer's rule to solve the following systems of equations:

(*i*) $x_1 - x_2 - x_3 = 1$
$2x_1 - 3x_2 + x_3 = 1$
$3x_1 + x_2 - x_3 = 2$

(*ii*) $x_1 + x_2 + x_3 = 6$
$x_1 + 2x_2 + 3x_3 = 14$
$x_1 - 2x_2 + x_3 = 2$

NOTES

2. Using the matrix inversion method to solve the following systems of equation:

   (*i*) $4x_1 - x_2 + 2x_3 = 15$
   $x_1 - 2x_2 - 3x_3 = -5$
   $5x_1 - 7x_2 + 9x_3 = 8$

   (*ii*) $x_1 + 4x_2 + 9x_3 = 16$
   $2x_1 + x_2 + x_3 = 10$
   $3x_1 + 2x_2 + 3x_3 = 18$

3. Solve the following systems of equation using Gaussian elimination method:

   (*i*) $2x + 2y + 4z = 18$
   $x + 3y + 2x = 13$
   $3x + y + 3x = 14$

   (*ii*) $x_1 + 2x_2 + x_3 + 4x_4 = 13$
   $x_1 + 4x_3 + 3x_4 = 28$
   $4x_1 + 2x_2 + 2x_3 + x_4 = 20$
   $-3x_1 + x_2 + 3x_3 + 2x_4 = 6$

4. Apply Gauss-Jordan elimination method to solve the following systems:

   (*i*) $x_1 + 2x_2 + 3x_3 = 4$
   $x_1 + x_2 + x_3 = 3$
   $2x_1 + 2x_2 + x_3 = 1$

   (*ii*) $5x_1 + 3x_2 + x_3 = 2$
   $4x_1 + 10x_2 + 4x_3 = -4$
   $2x_1 + 3x_2 + 5x_3 = 11$

5. Compute the solution of the following systems correct to three significant digits using Gauss-Jordan iteration method:

   (*i*) $9x_1 - 3x_2 + 2x_3 = 23$
   $6x_1 + 3x_2 + 14x_3 = 38$
   $4x_1 + 2x_2 - 3x_3 = 35$

   (*ii*) $x_1 + 2x_2 + 3x_3 + 4x_4 = 30$
   $4x_1 + x_2 + 2x_3 + 3x_4 = 24$
   $3x_1 + 4x_2 + x_3 + 2x_4 = 22$
   $2x_1 + 3x_2 + 4x_3 + x_4 = 24$

## 4.7 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

# BLOCK - II
# EIGEN VECTORS, INTERPOLATION, APPROXIMATION, DIFFERENTIATION AND INTEGRATION

# UNIT 5 EIGEN VALUES AND EIGEN VECTORS

**Structure**



## 5.0 INTRODUCTION

In linear algebra, an eigenvector or characteristic vector of a linear transformation is a nonzero vector that changes at most by a scalar factor when that linear transformation is applied to it. The corresponding eigenvalue is the factor by which the eigenvector is scaled. Geometrically, an eigenvector, corresponding to a real nonzero eigenvalue, points in a direction in which it is stretched by the transformation and the eigenvalue is the factor by which it is stretched. If the eigenvalue is negative, the direction is reversed. Loosely speaking, in a multidimensional vector space, the eigenvector is not rotated. However, in a one-dimensional vector space, the concept of rotation is meaningless.

In this unit, you will study about the Eigen values, Eigen vectors and Jacobi power method.

## 5.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the concept of Eigen values and Eigen vectors
- Analyse the Jacobi and power methods

## 5.2 FINDING EIGEN VALUES AND EIGEN VECTORS

NOTES

Let $A = [a_{ij}]$ be a square matrix of order $n$. If there exists a non-zero (non-null) column vector $X$ and a scalar $l$ such that,

$$AX = \lambda X$$

Then $\lambda$ is called an eigenvalue of the matrix $A$ and $X$ is called eigenvectors corresponding to the eigenvalue $\lambda$.

The problem of finding the values of the parameter $\lambda$, for which the homogeneous system,

$$AX = \lambda X \qquad ...(5.1)$$

possesses non-trivial solution is known as characteristic value problem or eigenvalue problem.

Thus, system of Equation (5.1) possesses non-trivial solution if and only if,

$$[A - \lambda I] = 0 \qquad ...(5.2)$$

This Equation is known as the characteristic equation of the matrix $A$.

The roots of this Equation (5.2) are called latent roots or characterstics values or eigenvalues of the matrix $A$. The corresponding non-trivial solutions are called eigenvectors or characteristic vectors of $A$.

If $A$ is an $n \times n$ matrix, then its characteristic equation is an $n$th degree polynomial equations in $\lambda$. Therefore, an $n \times n$ matrix has $n$ eigenvalues (real or complex).

Suppose $\lambda_i$ $(i = 1, 2, 3, \ldots, n)$ be the eigenvalues of $A$, then for each $\lambda_i$, there exists a non-null vector $X_i$ such that

$$AX_1 = \lambda_i X_i \qquad (i = 1, 2, 3, \ldots\ldots, n)$$

Multiplying both sides by a non-zero scalar $k$, we get

$$A(kX_i) = \lambda_i (kX_i)$$

This implies that an eigenvector is determined upto a multiplicative scalar. In other words, the eigenvector is not unique. But corresponding to an eigenvector of the matrix $A$, there can be one and only one eigenvalue of the matrix $A$.

It can be shown that for a matrix $A$ of order $n$, the characteristic Equation (5.2) can be written as,

$$\lambda^n - \beta_1\lambda^{n-1} + \beta_2\lambda^{n-2} + \ldots\ldots + (-1)^n \beta_n = 0$$

NOTES

Where $\beta_r$ is the sum of all the determinants formed from square matrices of order $r$ whose principal diagonals lie along the principal diagonal of $A$.

***Notes:***

1. An eigenvector of a matrix cannot correspond to two different eigenvalues.
2. An eigenvalue of a matrix can, and will correspond to different eigenvectors.

**Example 1:** Find the eigenvalues and eigenvectors of $A = \begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix}$

**Solution:** The characteristic equation is $\begin{bmatrix} 3-\lambda & 4 \\ 4 & -3-\lambda \end{bmatrix} = 0$

$$\lambda^2 - 25 = 0 \Rightarrow \lambda = \pm 5$$

The eigenvectors are given by $AX = \lambda X$

i.e., $$(A - \lambda I)X = 0$$

$$\begin{bmatrix} 3-\lambda & 4 \\ 4 & -3-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$(3 - \lambda)x_1 + 4x_2 = 0$$

$$4x_1 - (3 + \lambda)x_2 = 0$$

If, $\lambda = 5$,

We get, $-2x_1 + 4x_2 = 0$

$4x_1 - 8x_2 = 0$

Or, $$\frac{x_1}{2} = \frac{x_2}{1}$$

$\therefore$ The eigenvector is $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$

If $\lambda = -5$, we get: $8x_1 + 4x_2 = 0$

$4x_1 + 2x_2 = 0$

$2x_1 = -x_2$

$$\therefore \quad \frac{x_1}{-1} = \frac{x_2}{2}$$

$\therefore$ The eigenvector is $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$

The eigenvalues are 5 and –5 with the corresponding eigenvectors are $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$ respectively.

NOTES

**Example 2:** Find the eigenvalues and eigenvectors of the matrix $\begin{bmatrix} 3 & -4 & 4 \\ 1 & -2 & 4 \\ 1 & -1 & 3 \end{bmatrix}$

**Solution:** Let $A = \begin{bmatrix} 3 & -4 & 4 \\ 1 & -2 & 4 \\ 1 & -1 & 3 \end{bmatrix}$

The characteristic equation is $|A - \lambda I| = 0$

$$\text{i.e.,} \quad \begin{bmatrix} 3-\lambda & -4 & 4 \\ 1 & -2-\lambda & 4 \\ 1 & -1 & 3-\lambda \end{bmatrix} = 0$$

$$(3 - \lambda)[(-2 - \lambda)(3 - \sigma\lambda) + 4] + 4(3 - \lambda - 4) + 4(-1 + 2 + \lambda) = 0$$

$$\lambda^3 - 4\lambda^2 + \lambda + 6 = 0$$

The eigenvalues are the roots of this equation and they are –1, 2, 3.

The eigenvectors are given by the solution of:

$$(A - \lambda I)X = 0$$

$$\text{i.e.,} \begin{pmatrix} 3-\lambda & -4 & 4 \\ 1 & -2-\lambda & 4 \\ 1 & -1 & 3-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

$$\begin{aligned} (3 - \lambda)x_1 - 4x_2 + 4x_3 &= 0 \\ x_1 - (2 + \lambda)x_2 + 4x_3 &= 0 \\ x_1 - x_2 + (3 - \lambda)x_3 &= 0 \end{aligned}$$

**Case 1:** $\lambda = -1$ gives:

$$4x_1 - 4x_2 + 4x_3 = 0,\ x_1 - \boldsymbol{x}_2 + 4x_3 = 0,\ x_1 - x_2 + 4x_3 = 0$$

Solving the first and second equations by the method of cross multiplication we get:

$$\frac{x_1}{-12} = \frac{x_2}{-12} = \frac{x_3}{0}$$

$\therefore$ The eigenvector is $X_1 = \begin{bmatrix} -12 \\ -12 \\ 0 \end{bmatrix}$ or $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$

NOTES

**Case 2:** $\lambda = 2$ gives:

$$x_1 - 4x_2 + 4x_3 = 0,\ x_1 - 4x_2 + 4x_3 = 0,\ x_1 - x_2 + x_3 = 0$$

Solving the first and third equations we get the eigenvector $X_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$

**Case 3:** $\lambda = 3$ gives:

$$-4x_2 + 4x_3 = 0,\ x_1 - 5x_2 + 4x_3 = 0,\ x_1 - x_2 = 0$$

Solving any two of these equations we get the eigenvector $X_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

***Note:*** For a square matrix $A = (a_{ij})$ of order 3, the characteristic equation $|A - \lambda I| = 0$, takes the form:

$$\lambda^3 - \beta_1\lambda^2 + \beta_2\lambda - \beta_3 = 0$$

Where, $\beta_1 = a_{11} + a_{22} + a_{33}$ = Sum of the leading diagonal elements of $A$

$$\beta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}$$

= Sum of the minors of the leading diagonal elements of $A$

$\beta_3 = |A|$ = Determinant of the matrix $A$

In the Example 2

$$\beta_1 = 3 - 2 + 3 = 4$$

$$\beta_2 = \begin{vmatrix} -2 & 4 \\ -1 & 3 \end{vmatrix} + \begin{vmatrix} 3 & 4 \\ 1 & 3 \end{vmatrix} + \begin{vmatrix} 3 & -4 \\ 1 & -2 \end{vmatrix} = 1, \quad \beta_3 = -6$$

$\therefore$ The characteristic equation is $\lambda^3 - 4\lambda^2 + \lambda + 6 = 0$

***Note:*** If $\lambda$ is an eigenvalue of matrix $A$ of order 3, then the components of the eigenvector of matrix $A$ corresponding of $\lambda$ are proportional to the cofactors of the elements of any row of $|A - \lambda I|$ provided that not all of them vanish.

This method is used for the computation of the eigenvectors in the following examples.

**Example 3:** Find the eigenvalues and eigenvectors of the matrix $\begin{bmatrix} 2 & 2 & 0 \\ 2 & 1 & 1 \\ -7 & 2 & -3 \end{bmatrix}$.

**Solution:** The characteristic equation is,

$$\lambda^3 - \beta_1\lambda^2 + \beta_2\lambda - \beta_3 = 0,$$

Where, $\beta_1 = 1 + 2 - 3 = 0$, $\beta_2 = -5 - 6 - 2 = -13$, $\beta_3 = -12$

NOTES

∴ The characteristic equation is,

$$\lambda^3 - 13\lambda + 12 = 0$$

$$(\lambda - 1)(\lambda + 4)(\lambda - 3) = 0$$

∴ The eigenvalues are 1, –4, 3, when

$$\lambda = 1, A - \lambda I = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 0 & 1 \\ -7 & 2 & -4 \end{bmatrix}$$

Components of the eigenvectors are proportional to the cofactors of the elements of the first row, namely –2, 1, 4.

∴ The eigenvector is $\begin{bmatrix} -2 \\ 1 \\ 4 \end{bmatrix}$

$$\text{When } \lambda = 3, A - \lambda I = \begin{bmatrix} -1 & 2 & 0 \\ 2 & -2 & 1 \\ -7 & 2 & -6 \end{bmatrix}$$

The cofactors of the elements of the first row are 10, 5, –10, which are proportional to 2, 1, –2 respectively.

∴ The eigenvector is $\begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}$

$$\text{When } \lambda = -4, A - \lambda I = \begin{bmatrix} 6 & 2 & 0 \\ 2 & 5 & 1 \\ -7 & 2 & 1 \end{bmatrix}$$

The cofactors of the elements of the first row are 3, –9, 39, which are proportional to 1, –3, 13 respectively.

∴ The eigenvector is $\begin{bmatrix} 1 \\ -3 \\ 13 \end{bmatrix}$

**Example 4:** Find the eigenvalues and eigenvectors of $A = \begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

**NOTES**

**Solution:** The characteristic equation is $\lambda^3 - \beta_1\lambda^2 + \beta_2\lambda - \beta_3 = 0$

Where, $\beta_1 = 1 + 2 + 3 = 6$

$$\beta_2 = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 6 \\ 1 & 2 \end{bmatrix}$$

$$= 6 + 3 + 2 - 6 = 5$$

$$\beta_3 = |A| = 1(6) - 6(3) + 1(10) = -12$$

$\therefore$ The characteristic equation is $\lambda^3 - 6\lambda^2 + 5\lambda + 12 = 0$

The roots of this equation $-1, 3, 4$ are the eigenvalues of $A$, when

$$\lambda = -1, A - \lambda I = \begin{bmatrix} 2 & 6 & 1 \\ 1 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

The cofactors of the elements of the first row give the eigenvector as,

$$X_1 = \begin{bmatrix} 12 \\ -4 \\ 0 \end{bmatrix} \text{ or } \begin{bmatrix} 3 \\ -1 \\ 0 \end{bmatrix}$$

When $\lambda = 3, A - \lambda I = \begin{bmatrix} -2 & 6 & 1 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

Since, the cofactors of the elements of the 1st and 2nd rows vanish completely, we consider the cofactors of the elements of the 3rd row to get the eigenvector as,

$$X_2 = \begin{bmatrix} 1 \\ 1 \\ -4 \end{bmatrix}$$

When $\lambda = 4, A - \lambda I = \begin{bmatrix} -3 & 6 & 1 \\ 1 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}$

Considering the cofactors of the elements of the 1st row we get the eigenvector as,

$$X_3 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

NOTES

**Properties of Eigenvalues and Eigenvectors**

1. If all the eigenvalues of a matrix are distinct, then the corresponding eigenvectors are linearly independent.
2. If two or more eigenvalues of a matrix are equal then the corresponding eigenvectors may be linearly independent or linearly dependent.
3. The eigenvalues of a matrix and its transpose are the same. The characteristic equation of $A$ and $A^T$ (the transpose of $A$) are,

$$|A - \lambda I| = 0 \quad ...(5.3)$$

and $$|A^T - \lambda I| = 0 \quad ...(5.4)$$

LHS of Equation (5.4) is the determinant obtained by interchanging rows into columns of $|A - \lambda I|$. Since, the value of a determinant is unaltered by the interchanging of rows and columns, Equations (5.3) and (5.4) are identical. Therefore, the eigenvalues of a matrix and its transpose are the same.

4. The sum of the eigenvalues of a matrix $A$, is equal to the sum of the diagonal elements of $A$. The sum of the diagonal elements is called the *Trace* of the matrix $A$. The characteristic Equation of $A$ is,

$$\lambda^n - \beta_1\lambda^{n-1} + \beta_2\lambda^{n-2} - .... + (-1)^n\beta_n = 0 \quad ...(5.5)$$

Where, $\beta_1$ = sum of the diagonal elements of $A$. ...(5.6)

Let, $\lambda_1, \lambda_2,....., \lambda_n$ be the roots of Equation (5.5)

Then, $$\lambda_1 + \lambda_2 + .... + \lambda_n = -\frac{-\beta_1}{1} = \beta_1 \quad ...(5.7)$$

From Equations (5.6) and (5.7) we find that the sum of the eigenvalues is equal to the sum of the diagonal elements.

5. The product of the eigenvalues of a matrix $A$ is $|A|$.

The characteristic Equation of $A$ is,

$$\lambda_n - \beta_1\lambda^{n-1} + \beta_2\lambda^{n-2} - .... + (-1)^n\beta_n = 0 \quad ...(5.8)$$

Where, $\beta_n$ = Determinant of $A$.

If, $\lambda_1, \lambda_2,....,\lambda_n$ be the roots of Equation (5.8) then

$$\lambda_1, \lambda_2,...,\lambda_n = (-1)^n(-1)^n\frac{\beta_n}{1} = \beta_n \quad ...(5.9)$$

From Equations (5.8) and (5.9) we find that the product of the eigenvalues is equal to the value of the determinant of $A$.

6. The eigenvalues of a triangular matrix are the diagonal elements of it.

NOTES

***Notes:***

1. The sum of the eigenvalues of a matrix $A$, is equal to the sum of the diagonal elements of $A$, which is called the *Trace* of the matrix $A$.
2. If one of the eigenvalues is zero, then the matrix is singular and conversely, when the matrix is singular then at least one of the eigenvalues ought to be zero.
3. The eigenvalues of a diagonal matrix are the diagonal elements of it.

7. If $\lambda_i$ for($i = 1,2,3,...,n$) are the eigenvalues of $A$, then:

(*i*) $k\lambda_i$ for ($i = 1,2,3,....,n$) are the eigenvalues of the matrix $kA$, $k$ being a non-zero scalar.

(*ii*) $\frac{1}{\lambda_i}$, ($i = 1,2,3,....,n$) are the eigenvalues of the inverse matrix $A^{-1}$, provided $\lambda_i \neq 0$

(*i*) Let $X_i$ ($i = 1,2,3,....,n$) be the eigenvectors of the matrix $A$ corresponding to the eigenvalues of $\lambda_i$($i = 1,2,3,....,n$). Then,

$$AX_i = \lambda_i X_i \quad (i = 1,2,3,....,n) \qquad ...(5.10)$$

Multiplying by $k$, (a non-zero scalar):

$$kAX_i = k\lambda_i X_i$$

This implies that $k\lambda_i$ for ($i = 1,2,3,....,n$) are the eigenvalues of $kA$.

(*ii*) Premultiply Equation (5.10) by $A^{-1}$

$$A^{-1}AX_i = A^{-1}\lambda_i X_i$$

$$IX_i = \lambda_i A^{-1}X_i \text{ or } A^{-1}X_i = \lambda_i^{-1}X_i$$

This implies that $\lambda_i^{-1}$, for ($i = 1,2,3,...,n$) are the eigenvalues of $A^{-1}$

In general, if $\lambda_i$($i = 1,2,3,...,n$) are the eigenvalues of $A$, then $\lambda_i^m$($i = 1,2,3,...,n$), where $m$ is an integer, are the eigenvalues of $A^m$.

***Note:*** $A$ and $A^m$ ( $m$ being an integer) have the same eigenvectors even though the eigenvalues are different.

**Example 5:** Find the sum of the squares of the eigenvalues of $\begin{bmatrix} 3 & 0 & 0 \\ 8 & 4 & 0 \\ 6 & 2 & 5 \end{bmatrix}$.

**Solution:** The eigenvalues are 3, 4 and 5. Hence, the sum of the squares of eigenvalues =50.

**Example 6:** Two eigenvalues of matrix $\begin{bmatrix} 6 & -2 & 2 \\ -2 & 3 & -1 \\ 2 & -1 & 3 \end{bmatrix}$ are 2 and 8. Find the third eigenvalue.

**Solution:** Sum of the eigenvalues = Sum of the diagonal elements = 6 + 3 + 3 = 12.

Since the sum of the 2 given eigenvalues is 10 (2+8), the third eigenvalues is 12 – 10 = 2.

**Example 7:** If 3 and 15 are the two eigenvalues of $\begin{bmatrix} 8 & -6 & 2 \\ -6 & 7 & -4 \\ 2 & -4 & 3 \end{bmatrix}$, find the value of the determinant.

**Solution:** Let, $\lambda_1, \lambda_2, \lambda_3$ be the eigenvalues.

Then, $\lambda_1 + \lambda_2 + \lambda_3 = 8 + 7 + 3$; $3 + 15 + \lambda_3 = 18 \therefore \lambda_3 = 0$

The value of the determinant = Product of the eigenvalues = 0

$\therefore$ Values of the determinant is zero.

**Example 8:** If one of the eigenvalues of a matrix is zero, then what is the type of matrix?

**Solution:** The matrix is singular.

**Example 9:** Find the eigenvalues and eigenvectors of $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$.

**Solution:** The characteristic equation is $\lambda^3 - 3\lambda - 2 = 0$. Solving this equation we get the eigenvalues as –1, –1 and 2.

The eigenvalues are given by $\begin{bmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$

$$-\lambda x_1 + x_2 + x_3 = 0, x_1 - \lambda x_2 + x_3 = 0, x_1 + x_2 - \lambda x_3 = 0$$

**Case 1:** $\lambda = 2$ gives:

$$-2x_1 + x_2 + x_3 = 0, x_1 - 2x_2 + x_3 = 0, x_1 + x_2 - 2x_3 = 0$$

Solving any two of these equations we get eigenvector $X_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

**NOTES**

**Case 2:** $\lambda = -1$ gives:

$$x_1 + x_2 + x_3 = 0,\ x_1 + x_2 + x_3 = 0,\ x_1 + x_2 + x_3 = 0$$

Solving any two of these equations we get $x_1 = 0, x_2 = 0, x_3 = 0$ and the vector $x_2$ becomes a null vector which cannot be an eigenvector. This is because of the fact that all the three equations are one and the same. The rank of coefficient matrix is 1. Therefore, the system will have $(n - r) = (3 - 1) = 2$ linearly independent solutions. This indicates that, corresponding to $\lambda = -1$, there will be two linearly independent eigenvectors.

To get the solutions, we assign arbitrary values to two of the three variables as shown below. Considering the equation $x_1 + x_2 + x_3 = 0$ and assigning $x_3 = 0, x_2 = 1$, we get $x_1 = -1$.

$\therefore$ The eigenvector is $X_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$

Similarly, assigning the value $x_1 = 0, x_2 = 1$ we get $x_3 = -1$, so that the eigenvector is,

$$X_3 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

**Example 10:** Show that the eigenvalues of $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ are $-1$, 3 and verify that the eigenvalues are 1 and 9 for $A^2$.

**Solution:** The characteristic equation is $\lambda^2 - 2\lambda - 3 = 0$. The eigenvalues are $-1$, 3 and the corresponding eigenvectors are $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. By property, the eigenvectors of $A^2$ are 1, 9 and the corresponding eigenvectors are $(-1, 1)^T$, $(1, 1)^T$

**Verification:** $A^2 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$ has the characteristic equation, $\lambda^2 - 10\lambda + 9 = 0$.

This equation gives the eigenvalues 1, 9 and eigenvectors $\begin{pmatrix}-1\\1\end{pmatrix}$ and $\begin{pmatrix}1\\1\end{pmatrix}$.

NOTES

### Inner Product

Inner product or scalar product of two vectors $X$ and $Y$, denoted by $< X, Y>$ is defined as the scalar $X^TY$

i.e., $\langle X, X\rangle = X^T Y = \sum_{i=1}^{n} x_i y_i$

Inner product $\langle X, X\rangle$ is known as the square of the length of the vector $X$ and it is denoted as $|X|^2$, where $|X|$ is read as norm $X$. If $|X| = 1$ then $X$ is called a unit vector.

If the inner product between two vectors vanishes, then we say that the two vectors are orthogonal to each other.

**Example 11:** Show that $X = (1, -1, 2)^T$ and $Y = (3, -1, -2)^T$ are orthogonal.

**Solution:** $< X, Y > = X^TY = (1, -1, 2)\begin{pmatrix}3\\-1\\-2\end{pmatrix} = 3 + 1 = 4 = 0.$

Hence, $X$ and $Y$ are orthogonal.

### Eigenvalues of Real Symmetric Matrix

Let $\lambda$ be an eigenvalue and $X$ be the corresponding eigenvectors of the real symmetric matrix $A$. Then,

$$AX = \lambda X$$

Multiplying by $\bar{X}'$ on both sides,

$$\bar{X}'AX = \bar{X}'\lambda X \quad ...(5.11)$$

Taking complex conjugate on both sides,

$$\overline{\bar{X}'AX} = \overline{\bar{X}'\lambda X}$$

$$X'\bar{A}\bar{X} = X'\bar{\lambda}\bar{X}$$

Since $A$ is real, $\bar{A} = A$

$$X'A\bar{X} = X'\overline{\lambda X}$$

Taking transpose on both sides,

$$\left(X'A\bar{X}\right)' = \left(X'\overline{\lambda X}\right)'$$

$$\bar{X}'A'X = \bar{X}'\bar{\lambda}X = \overline{\lambda X}'X$$

Since $A$ is symmetric $A' = A$

NOTES

$$\therefore \qquad \bar{X}'AX = \overline{\lambda X}'X \qquad ...(5.12)$$

From Equations (5.11) and (5.12) we get

$$\lambda\bar{X}'X = \overline{\lambda X}'X$$

$$(\lambda - \bar{\lambda})\bar{X}'X = 0$$

Since $\bar{X}'X$ is non-zero, $\lambda = \bar{\lambda}$. Therefore $\lambda$ is real.

**Theorem 1:** *If $X_1$ and $X_2$ are two eigenvectors corresponding to two different eigenvalues $\lambda_1$ and $\lambda_2$ of a real symmetric matrix A, then $X_1$ and $X_2$ are orthogonal.*

**Proof:** Since $X_1$ and $X_2$ are the eigenvectors of matrix $A$ corresponding to the eigenvalues $\lambda_1$and $\lambda_2$, we have,

$$AX_1 = \lambda_1 X_1 \qquad (i)$$

$$AX_2 = \lambda_2 X_2 \qquad (ii)$$

Premultiplying Equation (*i*) by $X_2'$ we get,

$$X_2' AX_1 = X'_2\lambda_1 X_1$$

Taking transpose on both sides we get,

$$(X_2' AX_1)' = (X'_2\lambda_1 X_1)'$$

$$X_1' A' X_2 = X_1' \lambda_1 X_2$$

$$X_1' AX_2 = \lambda_1 X_1' X_2 \quad \text{where } A' = A, \text{ since } A \text{ is symmetric.}$$

$$X_1' \lambda_2 X_2 = \lambda_1 X_1' X_2 \qquad \text{[Using Equation } (ii)]$$

$$(\lambda_2 - \lambda_1) X_1' X_2 = 0$$

Since, $(\lambda_2 \neq \lambda_1)$, $X_1' X_2 = 0$

i.e., $X_1$ and $X_2$ are orthogonal.

**Example 12:** Find the eigenvalues of the matrix $\begin{bmatrix} 10 & -2 & -5 \\ -2 & 2 & 3 \\ -5 & 3 & 5 \end{bmatrix}$ and verify that the eigenvectors are mutually orthogonal.

**Solution:** Let $A = \begin{bmatrix} 10 & -2 & -5 \\ -2 & 2 & 3 \\ -5 & 3 & 5 \end{bmatrix}$

**NOTES**

$\beta_1 = 17; \beta_2 = 42; \beta_3 = 0$

The characteristic equation is,

$$\lambda^3 - 17\lambda^2 + 42\lambda = 0$$

$$\lambda(\lambda^2 - 17\lambda + 42) = 0$$

$$(\lambda - 3)(\lambda - 14) = 0$$

The eigenvalues are 0, 3 and 14. To find the eigenvectors we consider,

$$\begin{bmatrix} 10-\lambda & -2 & -5 \\ -2 & 2-\lambda & 3 \\ -5 & 3 & 5-\lambda \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$(10-\lambda)x_1 - 2x_2 - 5x_3 = 0$$

$$-2x_1 + (2-\lambda)x_2 + 3x_3 = 0$$

$$-5x_1 + 3x_2 + (5-\lambda)x_3 = 0$$

**Case 1:** $\lambda = 0$ gives:

$$10x_1 - 2x_2 - 5x_3 = 0, -2x_1 + 2x_2 + 3x_3 = 0$$

Solving we get eigenvector $X_1 = \begin{bmatrix} 1 \\ -5 \\ 4 \end{bmatrix}$

**Case 2:** $\lambda = 3$ gives:

$$7x_1 - 2x_2 - 5x_3 = 0, -2x_1 - x_2 + 3x_3 = 0$$

Solving we get eigenvector $X_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

**Case 3:** $\lambda = 14$ gives:

$$-4x_1 - 2x_2 - 5x_3 = 0, -2x_1 + 12x_2 + 3x_3 = 0$$

Solving we get eigenvector $X_3 = \begin{bmatrix} -3 \\ 1 \\ 2 \end{bmatrix}$

$\therefore$ The eigenvectors are $\begin{bmatrix} 1 \\ -5 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ 1 \\ 2 \end{bmatrix}$

NOTES

$$X_1' X_2 = 0, X_2' X_3 = 0 \text{ and } X_3' X_1 = 0$$

Hence, the three eigenvectors are mutually orthogonal.

***Note:*** It may be seen that any matrix, whose elements are polynomials can be expressed as a polynomial whose coefficients are matrices and vice versa. The following two examples illustrate this concept.

**Example 13:** Express the following matrices as polynomials with matrix coefficients.

$$(i) \quad \begin{pmatrix} \lambda + 2\lambda^2 & \lambda^3 - 3 \\ 1 + 3\lambda & -\lambda^2 \end{pmatrix},$$

$$(ii) \quad \begin{pmatrix} 1 + \lambda + \lambda^2 & \lambda + \lambda^2 - \lambda^3 & \lambda^3 - 3\lambda^2 + 5\lambda + 1 \\ \lambda^3 - 3\lambda^2 - 1 & \lambda + \lambda^2 & 1 - 3\lambda^2 + 4\lambda^3 \\ \lambda + \lambda^3 - 1 & 0 & \lambda^3 + \lambda^2 + \lambda + 1 \end{pmatrix}$$

**Solution:** (*i*) $\begin{pmatrix} \lambda + 2\lambda^2 & \lambda^3 - 3 \\ 1 + 3\lambda & -\lambda^2 \end{pmatrix}$

$$= \lambda^3 \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \lambda^2 \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix} + \lambda \begin{pmatrix} 1 & 0 \\ 3 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -3 \\ 1 & 0 \end{pmatrix}$$

$= A\lambda^3 + B\lambda^2 + C\lambda + D$, where $A, B$ and $C$ are matrices.

$$(ii) \quad \begin{pmatrix} 1 + \lambda + \lambda^2 & \lambda + \lambda^2 - \lambda^3 & \lambda^3 - 3\lambda^2 + 5\lambda + 1 \\ \lambda^3 - 3\lambda^2 - 1 & \lambda + \lambda^2 & 1 - 3\lambda^2 + 4\lambda^3 \\ \lambda + \lambda^3 - 1 & 0 & \lambda^3 + \lambda^2 + \lambda + 1 \end{pmatrix}$$

$$= A\lambda^3 + B\lambda^2 + C\lambda + D$$

Where,

$$A = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & 4 \\ 1 & 0 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 1 & -3 \\ -3 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 1 & 5 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \qquad D = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}$$

## 5.3 JACOBI AND POWER METHODS

In numerical linear algebra, the Jacobi method is an iterative algorithm for determining the solutions of a strictly diagonally dominant system of linear equations.

Each diagonal element is solved for, and an approximate value is plugged in. The process is then iterated until it converges. This algorithm is a stripped-down version of the Jacobi transformation method of matrix diagonalization. The method is named after Carl Gustav Jacob Jacobi.

In mathematics, power iteration, also known as the power method, is an eigenvalue algorithm: given a diagonalizable matrix $A$, the algorithm will produce a number $\lambda$, which is the greatest (in absolute value) eigenvalue of $A$, and a nonzero vector $v$, which is a corresponding eigenvector of $\lambda$, that is, $Av = \lambda v$. The algorithm is also known as the Von Mises iteration.

Power iteration is a very simple algorithm, but it may converge slowly. The most time-consuming operation of the algorithm is the multiplication of matrix $A$ by a vector, so it is effective for a very large sparse matrix with appropriate implementation.

**Jacobi Method**

As per the Jacobi method,

Let,

$A\mathbf{x} = \mathbf{b}$ be a square system of $n$ linear equations.

Where,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Then $A$ can be decomposed into a diagonal component $D$, a lower triangular part $L$ and an upper triangular part $U$:

$$A = D + L + U$$

Where,

$$D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

And,

$$L + U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}.$$

The solution is then obtained iteratively by means of,

$$\mathbf{x}^{(k+1)} = D^{-1}(\mathbf{b} - (L + U)\mathbf{x}^{(k)})$$

NOTES

Where $\mathbf{x}^{(k)}$ is referred as the $k$th approximation or iteration of $\mathbf{x}$ and $\mathbf{x}^{(k+1)}$ is the next or $k + 1$ iteration of{\displaystyle \mathbf {x} } $\mathbf{x}$. The element-based formula is thus:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \ldots, n.$$

The computation of $x_i^{(k+1)}$ requires each element in $\mathbf{x}^{(k)}$ except itself. Unlike the Gauss–Seidel method, we cannot overwrite $x_i^{(k)}$ with $x_i^{(k+1)}$, as that value will be necessary for the rest of the computation. The minimum amount of storage is two vectors of size $n$.

**Power Method**

The power iteration algorithm starts with a vector $b_0$, which may be an approximation to the dominant eigenvector or a random vector. The method is described by the recurrence relation,

$$b_{k+1} = \frac{Ab_k}{\|Ab_k\|}$$

Consequently, at every iteration, the vector $b_k$ is multiplied by the matrix $A$ and normalized.

If we assume that $A$ has an eigenvalue that is strictly greater in magnitude than its other eigenvalues and the starting vector $b_0$ has a nonzero component in the direction of an eigenvector associated with the dominant eigenvalue, then a subsequence $(b_k)$ converges to an eigenvector associated with the dominant eigenvalue.

Without the two assumptions above, the sequence $(b_k)$ does not necessarily converge. In this sequence,

$$b_k = e^{i\phi_k} v_1 + r_k,$$

Where $v_1$ is an eigenvector associated with the dominant eigenvalue, and $\|r_k\| \rightarrow 0$.. The presence of the term $e^{i\phi_k}$ implies that $(b_k)$ does not converge unless $e^{i\phi_k} = 1$. As per the two assumptions listed above, the sequence $(\mu_k)$ is defined by,

$$\mu_k = \frac{b_k^* A b_k}{b_k^* b_k}$$

This converges to the dominant eigenvalue.

NOTES

**Check Your Progress**

1. Define the terms eigenvalue and eigenvector.
2. Define any two properties of eigenvalues and eigenvectors.
3. Define inner product.

## 5.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Let $A = [a_{ij}]$ be a square matrix of order $n$. If there exists a non-zero (non-null) column vector $X$ and a scalar $l$ such that,

$$AX = \lambda X$$

Then $\lambda$ is called an eigenvalue of the matrix $A$ and $X$ is called eigenvectors corresponding to the eigenvalue $\lambda$.

2. (*i*) If all the eigenvalues of a matrix are distinct, then the corresponding eigenvectors are linearly independent.

   (*ii*) If two or more eigenvalues of a matrix are equal then the corresponding eigenvectors may be linearly independent or linearly dependent.

3. Inner product or scalar product of two vectors $X$ and $Y$, denoted by $<X, Y>$ is defined as the scalar $X^T Y$

$$\text{i.e., } \langle X, X \rangle = X^T Y = \sum_{i=1}^{n} x_i y_i$$

Inner product $\langle X, X \rangle$ is known as the square of the length of the vector $X$ and it is denoted as $|X|^2$, where $|X|$ is read as norm $X$. If $|X| = 1$ then $X$ is called a unit vector.

## 5.5 SUMMARY

- Let $A = [a_{ij}]$ be a square matrix of order $n$. If there exists a non-zero (non-null) column vector $X$ and a scalar $l$ such that,

  $AX = \lambda X$

  Then $\lambda$ is called an eigenvalue of the matrix $A$ and $X$ is called eigenvectors corresponding to the eigenvalue $\lambda$.

NOTES

- If $A$ is an $n \times n$ matrix, then its characteristic equation is an $n$th degree polynomial equations in $\lambda$. Therefore, an $n \times n$ matrix has $n$ eigenvalues (real or complex).
- An eigenvector of a matrix cannot correspond to two different eigenvalues.
- An eigenvalue of a matrix can, and will correspond to different eigenvectors.
- If $\lambda$ is an eigenvalue of matrix $A$ of order 3, then the components of the eigenvector of matrix $A$ corresponding of $\lambda$ are proportional to the cofactors of the elements of any row of $|A - \lambda I|$ provided that not all of them vanish.
- If all the eigenvalues of a matrix are distinct, then the corresponding eigenvectors are linearly independent.
- If two or more eigenvalues of a matrix are equal then the corresponding eigenvectors may be linearly independent or linearly dependent.

## 5.6 KEY WORDS

- **Eigen Value Problem:** The problem of finding the values of the parameter $\lambda$, for which the homogeneous system, possesses non-trivial solution is known as characteristic value problem or eigenvalue problem.

  $AX = \lambda X$

- **Inner Product:** Inner product or scalar product of two vectors $X$ and $Y$, denoted by $<X, Y>$ is defined as the scalar $X^T Y$

  i.e., $\langle X, X \rangle = X^T Y = \sum_{i=1}^{n} x_i y_i$

## 5.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. Define any three properties of eigenvalues and eigenvectors.
2. What is an inner product?

### Long-Answer Questions

1. Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 6 & -2 & 2 \\ -2 & 3 & -1 \\ 2 & -1 & 3 \end{pmatrix}$.

NOTES

2. If $X_1$ and $X_2$ are eigenvectors corresponding to distinct eigenvalues of $\lambda_1$ and $\lambda_2$ of $A$, then show that $X_1$ and $X_2$ are linearly independent.
3. Find the eigenvalues of the following matrices:

   (*i*) $\begin{pmatrix} 5 & 2 \\ 2 & 3 \end{pmatrix}$ (*ii*) $\begin{pmatrix} 1 & 1+i \\ 1-i & 2 \end{pmatrix}$

   (*iii*) $\begin{pmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{pmatrix}$ (*iv*) $\begin{pmatrix} 3 & 10 & 5 \\ -2 & -3 & -4 \\ 3 & 5 & 7 \end{pmatrix}$

4. Let $k_1, k_2, k_3 > 0$ and A = $[a_{ij}]$ where $a_{ij} = \begin{cases} 1 & i = j \\ \dfrac{k_i}{k_j} & i \neq j \end{cases}$ $(i, j = 1, 2, 3)$

   Write the matrix $A$ and find its eigenvalues.
5. Write 3 matrices whose characteristics equation is $\lambda^2 - 7\lambda + 6 = 0$.
6. Find the eigenvalues and eigenvectors of $3 \times 3$ null matrix.
7. Find the sum and product of the eigenvalues of $\begin{bmatrix} 2 & 1 & -1 & 0 \\ 1 & 3 & 4 & 2 \\ -1 & 4 & 1 & 2 \\ 0 & 2 & 2 & 1 \end{bmatrix}$.
8. Eigenvalues of a matrix are 1, –1 and 2. Find the value of Trace ($A$) and determinant $A$.
9. $A = \begin{pmatrix} 7 & 4 & -4 \\ 4 & -8 & -1 \\ 4 & -1 & -8 \end{pmatrix}$. If one the eigenvalues of $A$ is –9, find the other two eigenvalues.

## 5.8 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

**NOTES**

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

NOTES

# UNIT 6 INTERPOLATION AND APPROXIMATION

**Structure**



## 6.0 INTRODUCTION

Interpolation is the process of defining a function that takes on specified values at specified points. Polynomial interpolation is the most known one-dimensional interpolation method. Its advantages lies in its simplicity of realization and the good quality of interpolants obtained from it. You will learn about the various interpolation methods, namely Lagrange's interpolation, Newton's forward and backward difference interpolation formulae, iterative linear interpolation and inverse interpolation.

In this unit, you will study about the interpolation, approximation, Hermite interpolation, piecewise and spline interpolation and bivariate interpolation.

## 6.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the interpolation and approximation
- Analyse the Hermite interpolation
- Explain the piecewise, spline interpolation and bivariate interpolation

## 6.2 INTERPOLATION AND APPROXIMATION

The problem of interpolation is very fundamental problem in numerical analysis. The term interpolation literally means reading between the lines. In numerical analysis, interpolation means computing the value of a function $f(x)$ in between values of $x$ in a table of values. It can be stated explicitly as 'given a set of $(n + 1)$ values $y_0$, $y_1, y_2, ..., y_n$ for $x = x_0, x_1, x_2, ..., x_n$ respectively. The problem of interpolation is to compute the value of the function $y = f(x)$ for some non-tabular value of $x$.'

NOTES

The computation is often made by finding a polynomial called interpolating polynomial of degree less than or equal to $n$ such that the value of the polynomial is equal to the value of the function at each of the tabulated points. Thus if,

$$\varphi(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{6.1}$$

is the interpolating polynomial of degree $\leq n$, then

$$\varphi(x_i) = y_i, \quad \text{for} \quad i = 0, 1, 2, ..., \; n \tag{6.2}$$

It is true that, in general, it is difficult to guess the type of function to approximate $f(x)$. In case of periodic functions, the approximation can be made by a finite series of trigonometric functions. Polynomial interpolation is a very useful method for functional approximation. The interpolating polynomial is also useful as a basis to develop methods for other problems such as numerical differentiation, numerical integration and solution of initial and boundary value problems associated with differential equations.

The following theorem, developed by Weierstrass, gives the justification for approximation of the unknown function by a polynomial.

**Theorem 6.1:** Every function which is continuous in an interval $(a, b)$ can be represented in that interval by a polynomial to any desired accuracy. In other words, it is possible to determine a polynomial $P(x)$ such that $|\; f(x) - P(x) \;| < \varepsilon$, for every $x$ in the interval $(a, b)$ where $\varepsilon$ is any prescribed small quantity. Geometrically, it may be interpreted that the graph of the polynomial $y = P(x)$ is confined to the region bounded by the curves $y = f(x) - \varepsilon$ and $y = f(x) + \varepsilon$ for all values of $x$ within $(a, b)$, however small $\varepsilon$ may be.

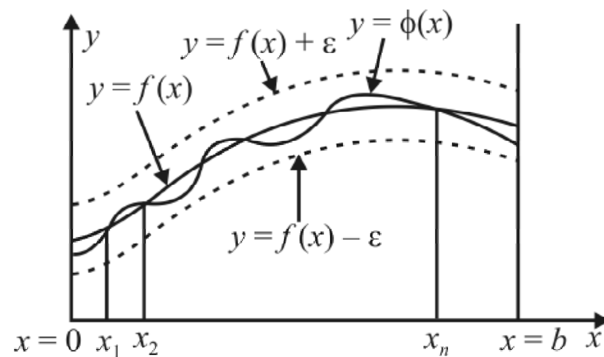



*Fig. 6.1* Interpolation

The following theorem is regarding the uniqueness of the interpolating polynomial.

NOTES

**Theorem 6.2:** For a real-valued function $f(x)$ defined at $(n+1)$ distinct points $x_0, x_1, ..., x_n$, there exists exactly one polynomial of degree $\le n$ which interpolates $f(x)$ at $x_0, x_1, ..., x_n$.

We know that a polynomial $P(x)$ which has $(n+1)$ distinct roots $x_0, x_1, ...,$ $x_n$ can be written as,

$$P(x) = (x - x_0)\ (x - x_1)\ .....(x - x_n)\ q\ (x)$$

where $q(x)$ is a polynomial whose degree is either 0 or $(n+1)$ which is less than the degree of $P(x)$.

Suppose that two polynomials $\varphi(x)$ and $\psi(x)$ are of degree $\le n$ and that both interpolate $f(x)$. Here $P(x) = \varphi(x) - \psi(x)$ at $x = x_0, x_1,..., x_n$. Then $P(x)$ vanishes at the $n+1$ points $x_0, x_1,..., x_n$. Thus $P(x) = 0$ and $\phi(x) = \psi(x)$.

**Iterative Linear Interpolation**

In this method, we successively generate interpolating polynomials, of any degree, by iteratively using linear interpolating functions.

Let $p_{01}(x)$ denote the linear interpolating polynomial for the tabulated values at $x_0$ and $x_1$. Thus, we can write as,

$$p_{01}(x) = \frac{(x_1 - x)f_0 - (x_0 - x)f_1}{x_1 - x_0}$$

This can be written with determinant notation as,

$$p_{01}(x) = \frac{\begin{vmatrix} f_0 & x_0 - x \\ f_1 & x_1 - x \end{vmatrix}}{x_1 - x_0} \qquad (6.3)$$

This form of $p_{01}(x)$ is easy to visualize and is convenient for desk computation. Thus, the linear interpolating polynomial through the pair of points $(x_0, f_0)$ and $(x_j, f_j)$ can be easily written as,

$$p_{0j}(x) = \frac{1}{x_j - x_0}\begin{vmatrix} f_0 & x_0 - x \\ f_j & x_j - x \end{vmatrix}, \text{ for } \quad j = 1, 2, ..., n \qquad (6.4)$$

Now, consider the polynomial denoted by $p_{01j}(x)$ and defined by,

$$p_{01j}(x) = \frac{1}{x_j - x_1}\begin{vmatrix} p_{01}(x) & x_1 - x \\ p_{0j}(x) & x_j - x \end{vmatrix}, \text{ for } \quad j = 2, 3, ..., n \qquad (6.5)$$

The polynomial $p_{01j}(x)$ interpolates $f(x)$ at the points $x_0, x_1, x_j$ $(j > 1)$ and is a polynomial of degree 2, which can be easily verified that,

$p_{0ij}(x_0) = f_0$, $p_{0ij}(x_i) = f_i$ and $p_{0ij}(x_j) = f_j$ because $p_{01}(x_0) = f_0 = p_{0ij}(x_0)$, etc.

Similarly, the polynomial $p_{012j}(x)$ can be constructed by replacing $p_{01}(x)$ by $p_{012}(x)$ and $p_{0j}(x)$ by $p_{01j}(x)$.

NOTES

Thus,

$$p_{012j}(x) = \frac{1}{x_j - x_2} \begin{vmatrix} p_{012}(x) & x_2 - x \\ p_{01j}(x) & x_j - x \end{vmatrix}, \text{ for } j = 3, 4, ..., n \quad (6.6)$$

Evidently, $p_{012j}(x)$ is a polynomial of degree 3 and it interpolates the function at $x_0$, $x_1$, $x_2$ and $x_j$.

i.e., $p_{012j}(x_0) = f_0;\ p_{012j}(x_i) = f_1;\ p_{012j}(x_2) = f_2$ and $p_{012j}(x_j) = f_j$

This process can be continued to generate higher and higher degree interpolating polynomials.

The results of the iterated linear interpolation can be conveniently represented as given in the following table.

| $x_k$ | $f_k$ | $p_{0j}$ | $p_{01j}$ | ... | $x_j - x$ |
|---|---|---|---|---|---|
| $x_0$ | $f_0$ | | | | $x_0 - x$ |
| $x_1$ | $f_1$ | $p_{01}$ | | | $x_1 - x$ |
| $x_2$ | $f_2$ | $p_{02}$ | $p_{012}$ | | $x_2 - x$ |
| $x_3$ | $f_3$ | $p_{03}$ | $p_{013}$ | | $x_3 - x$ |
| ... | ... | ... | ... | ... | ... |
| $x_j$ | $f_j$ | $p_{0j}$ | $p_{01j}$ | | $x_j - x$ |
| ... | ... | ... | ... | ... | ... |
| $x_n$ | $f_n$ | $p_{0n}$ | $p_{01n}$ | | $x_n - x$ |

The successive columns of interpolation results can be conveniently filled by computing the values of the determinants written using the previous column and the corresponding entries in the last column $x_j - x$. Thus, for computing $p_{01j}$'s for $j = 2, 3, ..., n$, we evaluate the determinant whose elements are the boldface quantities and divide the determinant's value by the difference $(x_j - x) - (x_1 - x)$.

**Example 1:** Find $s(2.12)$ using the following table by iterative linear interpolation:

| $x$ | 2.0 | 2.1 | 2.2 | 2.3 |
|---|---|---|---|---|
| $s(x)$ | 0.7909 | 0.7875 | 0.7796 | 0.7673 |

**Solution:** Here, $x = 2.12$. The following table gives the successive iterative linear interpolation results. The details of the calculations are shown below in the table.

| $x_j$ | $s(x_j)$ | $p_{0j}$ | $p_{01j}$ | $p_{012j}$ | $x_j - x$ |
|---|---|---|---|---|---|
| 2.0 | 0.7909 | | | | −0.12 |
| 2.1 | 0.7875 | **0.78682** | | | −0.02 |
| 2.2 | 0.7796 | 0.78412 | **0.78628** | | 0.08 |
| 2.3 | 0.7673 | 0.78146 | 0.78628 | **0.78628** | 0.18 |

$$p_{01} = \frac{1}{2.1-2.0}\begin{vmatrix} 0.7909 & -0.12 \\ 0.7875 & -0.02 \end{vmatrix} = 0.78682$$

$$p_{02} = \frac{1}{2.2-2.0}\begin{vmatrix} 0.7909 & -0.12 \\ 0.7796 & -0.08 \end{vmatrix} = 0.78412$$

$$p_{03} = \frac{1}{2.3-2.0}\begin{vmatrix} 0.7909 & -0.12 \\ 0.7673 & 0.18 \end{vmatrix} = 0.78146$$

$$p_{012} = \frac{1}{2.2-2.1}\begin{vmatrix} 0.78682 & -0.02 \\ 0.78412 & 0.08 \end{vmatrix} = 0.78628$$

$$p_{013} = \frac{1}{2.3-2.1}\begin{vmatrix} 0.78682 & -0.02 \\ 0.78146 & 0.18 \end{vmatrix} = 0.78628$$

$$p_{012} = \frac{1}{2.3-2.2}\begin{vmatrix} 0.78628 & 0.08 \\ 0.78628 & 0.18 \end{vmatrix} = 0.78628$$

The boldfaced results in the table give the value of the interpolation at $x =$ 2.12. The result 0.78682 is the value obtained by linear interpolation. The result 0.78628 is obtained by quadratic as well as by cubic interpolation. We conclude that there is no improvement in the third degree polynomial over that of the second degree.

***Notes*** **1.** Unlike Lagrange's methods, it is not necessary to find the degree of the interpolating polynomial to be used.

**2.** The approximation by a higher degree interpolating polynomial may not always lead to a better result. In fact it may be even worse in some cases.

Consider, the function $f(x) = 4$.

We form the finite difference table with values for $x = 0$ to 4.

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ | $\Delta^4 f(x)$ |
|---|---|---|---|---|---|
| 0 | 1 | | | | |
| | | 3 | | | |
| 1 | 4 | | 9 | | |
| | | 12 | | 27 | |
| 2 | 16 | | 36 | | 81 |
| | | 48 | | 108 | |
| 3 | 64 | | 144 | | |
| | | 192 | | | |
| 4 | 256 | | | | |

Newton's forward difference interpolating polynomial is given below by taking $x_0 = 0$,

$$u = \frac{x - x_0}{h} = x, \quad \varphi(x) = 1 + 3x + \frac{9}{2}x(x-1) + \frac{27}{6}x(x-1)(x-2) + \frac{81}{24}x(x-1)(x-2)(x-3)$$

NOTES

Now, consider values of $\varphi(x)$ at $x = 0.5$ by taking successively higher and higher degree polynomials.

Thus,

$$\varphi_1(0.5) = 1 + 0.5 \times 3 = 2.5, \quad \text{by linear interpolation}$$

$$\varphi_2(0.5) = 2.5 + \frac{0.5 \times (-0.5)}{2} \times 9 = 1.375, \quad \text{by quadratic interpolation}$$

$$\varphi_3(0.5) = 1.375 + \frac{0.5 \times (-0.5) \times (-1.5)}{6} \times 27 = 3.0625, \quad \text{by cubic interpolation}$$

$$\varphi_4(0.5) = 3.0625 + \frac{(0.5)(-0.5)(-1.5)(-2.5)}{24} \times 81 = -0.10156, \quad \text{by quartic interpolation}$$

We note that the actual value $4^{0.5} = 2$ is not obtainable by interpolation. The results for higher degree interpolating polynomials become worse.

***Note:*** Lagrange's interpolation formula and iterative linear interpolation can easily be implemented for computations by a digital computer.

**Example 2:** Determine the interpolating polynomial for the following table of data:

| | | | | |
|---|---|---|---|---|
| $x$ | 1 | 2 | 3 | 4 |
| $y$ | −1 | −1 | 1 | 5 |

**Solution:** The data is equally spaced. We thus form the finite difference table.

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ |
|---|---|---|---|
| 1 | −1 | | |
| | | 0 | |
| 2 | −1 | | 2 |
| | | 2 | |
| 3 | 1 | | 2 |
| | | 4 | |
| 4 | 5 | | |

Since the differences of second order are constant, the interpolating polynomial is of degree two. Using Newton's forward difference interpolation, we get

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0,$$

Here, $x_0 = 1, \quad u = x - 1.$

Thus, $y = -1 + (x-1) \times 0 + \dfrac{(x-1)(x-2)}{2} \times 2 = x^2 - 3x + 1.$

NOTES

**Example 3:** Compute the value of $f(7.5)$ by using suitable interpolation on the following table of data.

| $x$ | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| $f(x)$ | 28 | 65 | 126 | 217 | 344 | 513 |

**Solution:** The data is equally spaced. Thus for computing $f(7.5)$, we use Newton's backward difference interpolation. For this, we first form the finite difference table as shown below.

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ |
|---|---|---|---|---|
| 3 | 28 | | | |
| | | 37 | | |
| 4 | 65 | | 24 | |
| | | 61 | | 6 |
| 5 | 126 | | 30 | |
| | | 91 | | 6 |
| 6 | 217 | | 36 | |
| | | 127 | | 6 |
| 7 | 344 | | 42 | |
| | | 169 | | |
| 8 | 513 | | | |

The differences of order three are constant and hence we use Newton's backward difference interpolating polynomial of degree three.

$$f(x) = y_n + v \ \nabla y_n + \frac{v(v+1)}{2\ !}\nabla^2 y_n + \frac{v(v+1)(v+2)}{3\ !}\nabla^3 y_n,$$

$$v = \frac{x - x_n}{h}, \text{ for } x = 7.5, \quad x_n = 8$$

$$\therefore \qquad v = \frac{7.5 - 8}{1} = -0.5$$

$$f(7.5) = 513 - 0.5 \times 169 + \frac{(-0.5)\ \ (-0.5+1)}{2} \times 42 + \frac{-0.5 \times 0.5 \times 1.5}{6} \times 6$$

$$= 513 - 84.5 - 5.25 - 0.375$$

$$= 422.875$$

**Example 4:** Determine the interpolating polynomial for the following data:

| $x$ | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| $f(x)$ | 5 | 10 | 17 | 29 | 50 |

NOTES

**Solution:** The data is equally spaced. We construct the Newton's forward difference interpolating polynomial. The finite difference table is,

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ | $\Delta^4 f(x)$ |
|---|---|---|---|---|---|
| 2 | 5 | | | | |
| | | 5 | | | |
| 4 | 10 | | 2 | | |
| | | 7 | | 3 | |
| 6 | 17 | | 5 | | 1 |
| | | 12 | | 4 | |
| 8 | 29 | | 9 | | |
| | | 21 | | | |
| 10 | 50 | | | | |

Here, $x_0 = 2$, $u = (x–x_0)/h = (x–2)/2$.

The interpolating polynomial is,

$$f(x) = f(x_0) + u\,\Delta f(x_0) + \frac{u(u-1)}{2!}\Delta^2 f(x_0) + ...$$

$$= 5 + \frac{x-2}{2}\times 5 + \frac{x-2}{2}\left(\frac{x-2}{2}-1\right)\frac{2}{2!} + \frac{x-2}{2}\left(\frac{x-2}{2}-1\right)\left(\frac{x-2}{2}-2\right)\frac{3}{3!}$$

$$+\frac{x-2}{2}\left(\frac{x-2}{2}-1\right)\left(\frac{x-2}{2}-2\right)\left(\frac{x-2}{2}-3\right)\frac{1}{4!}$$

$$= \frac{1}{384}(x^4 + 4x^3 - 52x^2 + 1040x)$$

**Example 5:** Find the interpolating polynomial which takes the following values:

$y(0) = 1, y(0.1) = 0.9975,\ y(0.2) = 0.9900, y(0.3) = 0.9980$. Hence compute $y\,(0.05)$.

**Solution:** The data values of $x$ are equally spaced we form the finite difference table,

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 0.0 | 1.0000 | | | |
| | | −25 | | |
| 0.1 | 0.9975 | | −50 | |
| | | −75 | | 25 |
| 0.2 | 0.9900 | | −25 | |
| | | −100 | | |
| 0.3 | 0.9800 | | | |

Here, $h = 0.1$. Choosing $x_0 = 0.0$, we have $s = \frac{x}{0.1} = 10x.$ Newton's forward difference interpolation formula is,

NOTES

$$y = y_0 + s\,\Delta y_0 + \frac{s(s-1)}{2!}\Delta^2 y_0 + \frac{s(s-1)(s-2)}{3!}\Delta^3 y_0$$

$$= 1 + 10x(-0.0025) + \frac{10x(10x-1)}{2!}(-0.0050) + \frac{10x(10x-1)(10x-2)}{6}\times 0.0025$$

$$= 1.0 - 0.25x - 0.25x^2 + 0.25x + \frac{2.5}{6}x^3 - \frac{300}{4}\times 0.0025x^2 + \frac{0.025}{6}x$$

$$= 1.0 + 0.004x - 0.375x^2 + 0.421x^3$$

$$y(0.05) = 1.0002$$

**Example 6:** Compute $f(0.23)$ and $f(0.29)$ by using suitable interpolation formula with the table of data given below.

| $x$ | 0.20 | 0.22 | 0.24 | 0.26 | 0.28 | 0.30 |
|---|---|---|---|---|---|---|
| $f(x)$ | 1.6596 | 1.6698 | 1.6804 | 1.6912 | 1.7024 | 1.7139 |

**Solution:** The data being equally spaced, we use Newton's forward difference interpolation for computing $f(0.23)$, and for computing $f(0.29)$, we use Newton's backward difference interpolation. We first form the finite difference table,

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ |
|---|---|---|---|
| 0.20 | 1.6596 | | |
| | | 102 | |
| 0.22 | 1.6698 | | 4 |
| | | 106 | |
| 0.24 | 1.6804 | | 2 |
| | | 108 | |
| 0.26 | 1.6912 | | 4 |
| | | 112 | |
| 0.28 | 1.7024 | | 3 |
| | | 115 | |
| 0.30 | 1.7139 | | |

We observe that differences of order higher than two would be irregular. Hence, we use second degree interpolating polynomial. For computing $f(0.23)$, we take

$x_0 = 0.22$ so that $u = \frac{x - x_0}{h} = \frac{0.23 - 0.22}{0.02} = 0.5.$

Using Newton's forward difference interpolation, we compute

$$f(0.23) = 1.6698 + 0.5 \times 0.0106 + \frac{(0.5)(0.5-1.0)}{2} \times 0.0002$$

$$= 1.6698 + 0.0053 - 0.000025$$

$$= 1.675075$$

$$\approx 1.6751$$

Again for computing $f(0.29)$, we take $x_n = 0.30$,

NOTES

so that $v = \frac{x - x_n}{n} = \frac{0.29 - 0.30}{0.02} = -0.5$

Using Newton's backward difference interpolation we evaluate,

$$f(0.29) = 1.7139 - 0.5 \times .0115 + \frac{(-0.5)(-0.5+1.0)}{2} \times 0.0003$$

$$= 1.7139 - 0.00575 - 0.00004$$

$$= 1.70811$$

$$\simeq 1.7081$$

**Example 7:** Compute values of $e^x$ at $x = 0.02$ and at $x = 0.38$ using suitable interpolation formula on the table of data given below.

| $x$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|---|
| $e^x$ | 1.0000 | 1.1052 | 1.2214 | 1.3499 | 1.4918 |

**Solution:** The data is equally spaced. We have to use Newton's forward difference interpolation formula for computing $e^x$ at $x = 0.02$, and for computing $e^x$ at $x = 0.38$, we have to use Newton's backward difference interpolation formula. We first form the finite difference table.

| $x$ | $y = e^x$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ |
|---|---|---|---|---|---|
| 0.0 | 1.0000 | | | | |
| | | 1052 | | | |
| 0.1 | 1.1052 | | 110 | | |
| | | 1162 | | 13 | |
| 0.2 | 1.2214 | | 123 | | −2 |
| | | 1285 | | 11 | |
| 0.3 | 1.3499 | | 134 | | |
| | | 1419 | | | |
| 0.4 | 1.4918 | | | | |

For computing $e^{0.02}$, we take $x_0 = 0$

$$\therefore \quad u = \frac{x - x_0}{h} = \frac{0.02 - 0.0}{0.1} = 0.2$$

By Newton's forward difference interpolation formula, we have

$$e^{0.02} = 1.0 + 0.2 \times 0.1052 + \frac{0.2(0.2-1)}{2} \times 0.0110 + \frac{0.2(0.2-1)(0.2-2)}{6} \times 0.0013$$

$$+ \frac{0.2 \quad (0.2-1)(0.2-2)(0.2-3)}{24} \times -0.0002$$

$$= 1.0 + .02104 - 0.00088 + 0.00006 + 0.00001$$

$$= 1.02023 \simeq 1.0202$$

NOTES

For computing $e^{0.38}$ we take $x_n = 0.4$. Thus, $v = \frac{0.38 - 0.4}{0.1} = -0.2$

By Newton's backward difference interpolation formula, we have

$$e^{0.38} = 1.4918 + (-0.2) \times 0.1419 + \frac{(-0.2)(-0.2+1)}{2} \times 0.0134$$
$$+ \frac{(-0.2)(-0.2+1)(-0.2+2)}{6} \times 0.0011 + \frac{-0.2(-0.2+1)(-0.2+2)(-0.2+3)}{24} \times (-0.0002)$$
$$= 1.4918 - 0.02838 - 0.00107 - 0.00005 - 0.00001$$
$$= 1.49287 - 0.02844$$
$$= 1.46443 \approx 1.4644$$

### Lagrange's Interpolation

Lagrange's interpolation is useful for unequally spaced tabulated values. Let $y = f(x)$ be a real valued function defined in an interval $(a, b)$ and let $y_0, y_1, ..., y_n$ be the $(n + 1)$ known values of $y$ at $x_0, x_1, ..., x_n$, respectively. The polynomial $\varphi(x)$, which interpolates $f(x)$, is of degree less than or equal to $n$. Thus,

$$\varphi(x_i) = y_i, \quad \text{for} \quad i = 0, 1, 2, ..., n \tag{6.7}$$

The polynomial $\varphi(x)$ is assumed to be of the form,

$$\varphi(x) = \sum_{i=0}^{n} l_i(x) y_i \tag{6.8}$$

where each $l_i(x)$ is a polynomial of degree $\leq n$ in $x$ and is called Lagrangian function.

Now, $\varphi(x)$ satisfies Equation (6.7) if each $l_i(x)$ satisfies,

$$l_i(x_j) = 0 \quad \text{when} \quad i \neq j$$
$$= 1 \quad \text{when} \quad i = j \tag{6.9}$$

Equation (6.9) suggests that $l_i(x)$ vanishes at the $(n+1)$ points $x_0, x_1, ... x_{i-1}, x_{i+1}, ..., x_n$. Thus, we can write,

$$l_i(x) = c_i (x - x_0)(x - x_1) ... (x - x_{i-1})(x - x_{i+1})...(x - x_n)$$

where $c_i$ is a constant given by $l_i(x_i) = 1$,

i.e., $c_i \ (x_i - x_0)\ (x_i - x_1)...(x_i - x_{i-1})\ (x_i - x_{i+1})...\ (x_i - x_n) = 1$

Thus, $l_i(x) = \frac{(x - x_0)(x - x_1)...(x - x_{i-1})(x - x_{i+1})...(x - x_n)}{(x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)}$ for $i = 0, 1, 2, ..., n$ (6.10)

NOTES

Equations (6.8) and (6.10) together give Lagrange's interpolating polynomial.

**Algorithm:** To compute $f(x)$ by Lagrange's interpolation.

**Step 1:** Read $n$ [$n$ being the number of values]

**Step 2:** Read values of $x_i, f_i$ for $i = 1, 2,..., n$.

**Step 3:** Set sum = 0, $i = 1$

**Step 4:** Read $x$ [$x$ being the interpolating point]

**Step 5:** Set $j = 1$, product = 1

**Step 6:** Check if $j \neq i$, product = product $\times (x - x_j)/(x_i - x_j)$ else go to Step 7

**Step 7:** Set $j = j + 1$

**Step 8:** Check if $j > n$, then go to Step 9 else go to Step 6

**Step 9:** Compute sum = sum + product $\times f_i$

**Step 10:** Set $i = i + 1$

**Step 11:** Check if $i > n$, then go to Step 12
else go to Step 5

**Step 12:** Write $x$, sum

**Example 8:** Compute $f(0.4)$ for the table below by Lagrange's interpolation.

| $x$ | 0.3 | 0.5 | 0.6 |
|---|---|---|---|
| $f(x)$ | 0.61 | 0.69 | 0.72 |

**Solution:** The Lagrange's interpolation formula gives,

$$f(0.4) = \frac{(0.4-0.5)(0.4-0.6)}{(0.3-0.5)(0.3-0.6)} \times 0.61 + \frac{(0.4-0.3)(0.4-0.6)}{(0.5-0.3)(0.5-0.6)} \times 0.69 + \frac{(0.4-0.3)(0.4-0.5)}{(0.6-0.3)(0.6-0.5)} \times 0.72$$
$$= 0.203 + 0.69 - 0.24 = 0.653 \simeq 0.65$$

Thus, $f(0.4) = 0.65$.

**Example 9:** Using Lagrange's formula, find the value of $f(0)$ from the table given below.

| $x$ | $-1$ | $-2$ | 2 | 4 |
|---|---|---|---|---|
| $f(x)$ | $-1$ | $-9$ | 11 | 69 |

**Solution:** Using Lagrange's interpolation formula, we find

$$f(0) = \left[\frac{(0+2)(0-2)(0-4)}{(-1+2)(-1-2)(-1-4)} \times (-1)\right] + \left[\frac{(0+1)(0-2)(0-4)}{(-2+1)(-2-2)(-2-4)} \times (-9)\right]$$
$$+ \left[\frac{(0+1)(0+2)(0-4)}{(2+1)(2+2)(2-4)} \times 11\right] + \left[\frac{(0+1)(0+2)(0-2)}{(4+1)(4+2)(4-2)} \times 69\right]$$
$$= -\frac{16}{15} + \frac{9}{3} + \frac{11}{3} - \frac{69}{15} = \frac{20}{3} - \frac{85}{15}$$
$$= \frac{20}{3} - \frac{17}{3} = 1$$

**NOTES**

**Example 10:** Determine the interpolating polynomial of degree three for the table given below.

| $x$ | $-1$ | $0$ | $1$ | $2$ |
|---|---|---|---|---|
| $f(x)$ | $1$ | $1$ | $1$ | $-3$ |

**Solution:** We have Lagrange's third degree interpolating polynomial as,

$$f(x) = \sum_{i=0}^{3} l_i(x) f(x_i)$$

where

$$l_0(x) = \frac{(x-0)(x-1)(x-2)}{(-1-0)(-1-1)(-1-2)} = -\frac{1}{6}x(x-1)(x-2)$$

$$l_1(x) = \frac{(x+1)(x-1)(x-2)}{(0+1)(0-1)(0-2)} = \frac{1}{2}(x+1)(x-1)(x-2)$$

$$l_2(x) = \frac{(x+1)(x-0)(x-2)}{(1+1)(1-0)(1-2)} = -\frac{1}{2}(x+1)x(x-2)$$

$$l_3(x) = \frac{(x+1)(x-0)(x-1)}{(2+1)(2-0)(2-1)} = \frac{1}{6}(x+1)\,x\,(x-2)$$

$$f(x) = -\frac{1}{6}x(x-1)(x-2)\times 1 + \frac{1}{2}(x+1)(x-1)(x-2)\times 1 - \frac{1}{2}(x+1)\,x\,(x-2)\times 1 + \frac{1}{6}(x+1)x(x-2)\times(-3)$$

$$= -\frac{1}{6}\,(4x^3 - 4x - 6)$$

$$= \frac{-1}{3}(2x^3 - 2x - 3)$$

**Example 11:** Evaluate the values of $f(2)$ and $f(6.3)$ using Lagrange's interpolation formula for the table of values given below.

| $x$ | 1.2 | 2.5 | 4 | 5.1 | 6 | 6.5 |
|---|---|---|---|---|---|---|
| $f(x)$ | 6.84 | 14.25 | 27 | 39.21 | 51 | 58.25 |

**Solution:** It is not advisable to use a higher degree interpolating polynomial. For evaluation of $f(2)$ we take a second degree polynomial using the values of $f(x)$ at the points $x_0 = 1.2$, $x_1 = 2.5$ and $x_2 = 4$.

Thus,

$$f(2) = l_0(2) \times 6.84 + l_1(2) \times 14.25 + l_2(2) \times 27$$

Where

$$l_0(2) = \frac{(2-2.5)(2-4)}{(1.2-2.5)(1.2-4)} = 0.275$$

$$l_1(2) = \frac{(2-1.2)(2-4)}{(2.5-1.2)(2.5-4)} = 0.821$$

$$l_2(2) = \frac{(2-1.2)(2-2.5)}{(4-1.2)(4-2.5)} = -0.095$$

NOTES

$$\therefore f(2) = 0.275 \times 6.84 + 0.821 \times 14.25 - 0.095 \times 27 = 11.015 \simeq 11.02$$

For evaluation of $f(6.3)$, we consider the values of $f(x)$ at $x_0 = 5.1, x_1 = 6.0, x_2 = 6.5$.

Thus, $f(6.3) = l_0(6.3) \times 39.21 + l_1(6.3) \times 51 + l_2(6.3) \times 58.25$

where

$$l_0(6.3) = \frac{(6.3-6.0)(6.3-6.5)}{(5.1-6.0)(5.1-6.5)} = -0.048$$

$$l_1(6.3) = \frac{(6.3-5.1)(6.3-6.5)}{(6-5.1)(6.0-6.5)} = 0.533$$

$$l_2(6.3) = \frac{(6.3-5.1)(6.3-6.0)}{(6.5-5.1)(6.5-6.0)} = 0.514$$

$$\therefore f(6.3) = -0.048 \times 39.21 + 0.533 \times 51 + 0.514 \times 58.25$$
$$= 55.241 \cong 55.24$$

Since, the computed result cannot be more accurate than the data, the final result is rounded-off to the same number of decimals as the data. In some cases, a higher degree interpolating polynomial may not lead to better results.

### Interpolation for Equally Spaced Tabular Values

For interpolation of an unknown function when the tabular values of the argument $x$ are equally spaced, we have two important interpolation formulae, viz.,

(*i*) Newton's forward difference interpolation formula

(*ii*) Newton's backward difference interpolation formula

We will first discuss the finite differences which are used in evaluating the above two formulae.

### Finite Differences

Let us assume that values of a function $y = f(x)$ are known for a set of equally spaced values of $x$ given by $\{x_0, x_1, ..., x_n\}$, such that the spacing between any two consecutive values is equal. Thus, $x_1 = x_0 + h$, $x_2 = x_1 + h, ..., x_n = x_{n-1} + h$, so that $x_i = x_0 + ih$ for $i = 1, 2, ..., n$. We consider two types of differences known as forward differences and backward differences of various orders. These differences can be tabulated in a finite difference table as explained in the subsequent sections.

### Forward Differences

Let $y_0, y_1, ..., y_n$ be the values of a function $y = f(x)$ at the equally spaced values of $x = x_0, x_1, ..., x_n$. The differences between two consecutive $y$ given by $y_1 - y_0, y_2 - y_1, ..., y_n - y_{n-1}$ are called the first order forward differences of the function $y = f(x)$ at the points $x_0, x_1, ..., x_{n-1}$. These differences are denoted by,

$$\Delta y_0 = y_1 - y_0, \quad \Delta y_1 = y_2 - y_1, \ ..., \Delta y_{n-1} = y_n - y_{n-1} \qquad (6.11)$$

NOTES

where $\Delta$ is termed as the forward difference operator defined by,

$$\Delta f(x) = f(x+h) - f(x) \qquad (6.12)$$

Thus, $\Delta y_i = y_{i+1} - y_i$, for $i = 0, 1, 2, ..., n-1$, are the first order forward differences at $x_i$.

The differences of these first order forward differences are called the second order forward differences.

Thus,

$$\Delta^2 y_i = \Delta(\Delta y_i)$$
$$= \Delta y_{i+1} - \Delta y_i, \text{ for } i = 0, 1, 2, ..., n-2 \qquad (6.13)$$

Evidently,

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = y_2 - y_1 - (y_1 - y_0) = y_2 - 2y_1 + y_0$$

And, $\Delta^2 y_i = y_{i+2} - y_{i+1} - (y_{i+1} - y_i)$

i.e., $\Delta^2 y_i = y_{i+2} - 2y_{i+1} + y_i, \text{ for } i = 0, 1, 2, ..., n-2 \qquad (6.14)$

Similarly, the third order forward differences are given by,

$$\Delta^3 y_i = \Delta^2 y_{i+1} - \Delta^2 y_i, \text{ for } i = 0, 1, 2, ..., n-3$$

i.e., $\Delta^3 y_i = y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i \qquad (6.15)$

Finally, we can define the *n*th order forward difference by,

$$\Delta^n y_0 = y_n - ny_{n-1} + \frac{n(n-1)}{2!} y_{n-2} + ... + (-1)^n y_0 \qquad (6.16)$$

The coefficients in above equations are the coefficients of the binomial expansion $(1-x)^n$.

The forward differences of various orders for a table of values of a function $y = f(x)$, are usually computed and represented in a diagonal difference table. A diagonal difference table for a table of values of $y = f(x)$, for six points $x_0, x_1, x_2, x_3, x_4, x_5$ is shown here.

NOTES

Diagonal difference Table for $y = f(x)$:

| $i$ | $x_i$ | $y_i$ | $\Delta y_i$ | $\Delta^2 y_i$ | $\Delta^3 y_i$ | $\Delta^4 y_i$ | $\Delta^5 y_i$ |
|---|---|---|---|---|---|---|---|
| 0 | $x_0$ | $y_0$ | | | | | |
| | | | $\Delta y_0$ | | | | |
| 1 | $x_1$ | $y_1$ | | $\Delta^2 y_0$ | | | |
| | | | $\Delta y_1$ | | $\Delta^3 y_0$ | | |
| 2 | $x_2$ | $y_2$ | | $\Delta^2 y_1$ | | $\Delta^4 y_0$ | |
| | | | $\Delta y_2$ | | $\Delta^3 y_1$ | | $\Delta^5 y_0$ |
| 3 | $x_3$ | $y_3$ | | $\Delta^2 y_2$ | | $\Delta^4 y_1$ | |
| | | | $\Delta y_3$ | | $\Delta^3 y_2$ | | |
| 4 | $x_4$ | $y_4$ | | $\Delta^2 y^3$ | | | |
| | | | $\Delta y_4$ | | | | |
| 5 | $x_5$ | $y_5$ | | | | | |

The entries in any column of the differences are computed as the differences of the entries of the previous column and one placed in between them. The upper data in a column is subtracted from the lower data to compute the forward differences. We notice that the forward differences of various orders with respect to $y_i$ are along the forward diagonal through it. Thus $\Delta y_0$, $\Delta^2 y_0$, $\Delta^3 y_0$, $\Delta^4 y_0$ and $\Delta^5 y_0$ lie along the top forward diagonal through $y_0$. Consider the following example.

**Example 12:** Given the table of values of $y = f(x)$,

| $x$ | 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|
| $y$ | 8 | 12 | 21 | 36 | 62 |

form the diagonal difference table and find the values of $\Delta f(5)$, $\Delta^2 f(3)$, $\Delta^3 f(1)$.

**Solution:** The diagonal difference table is,

| $i$ | $x_i$ | $y_i$ | $\Delta y_i$ | $\Delta^2 y_i$ | $\Delta^3 y_i$ | $\Delta^4 y_i$ |
|---|---|---|---|---|---|---|
| 0 | 1 | 8 | | | | |
| | | | 4 | | | |
| 1 | 3 | 12 | | 5 | | |
| | | | 9 | | 1 | |
| 2 | 5 | 21 | | 6 | | 4 |
| | | | 15 | | 5 | |
| 3 | 7 | 36 | | 11 | | |
| | | | 26 | | | |
| 4 | 9 | 62 | | | | |

From the table, we find that $\Delta f(5) = 15$, the entry along the diagonal through the entry 21 of $f(5)$.

Similarly, $\Delta^2 f(3) = 6$, the entry along the diagonal through $f(3)$. Finally, $\Delta^3 f(1) = 1$.

NOTES

### Backward Differences

The backward differences of various orders for a table of values of a function $y = f(x)$ are defined in a manner similar to the forward differences. The backward difference operator $\nabla$ (inverted triangle) is defined by $\nabla f(x) = f(x) - f(x-h)$.

Thus, $\nabla y_k = y_k - y_{k-1}$, for $k = 1, 2, ..., n$

i.e., $\nabla y_1 = y_1 - y_0, \nabla y_2 = y_2 - y_1, ..., \nabla y_n = y_n - y_{n-1}$

(6.17)

The backward differences of second order are defined by,

$$\nabla^2 y_k = \nabla y_k - \nabla y_{k-1} = y_k - 2y_{k-1} + y_{k-2}$$

Hence,

$$\nabla^2 y_2 = y_2 - 2y_1 + y_0, \text{ and } \nabla^2 y_n = y_n - 2y_{n-1} + y_{n-2} \quad (6.18)$$

Higher order backward differences can be defined in a similar manner.

Thus, $\nabla^3 y_n = y_n - 3y_{n-1} + 3y_{n-2} - y_{n-3}$, etc.

(6.19)

Finally,

$$\nabla^n y_n = y_n - ny_{n-1} + \frac{n(n-1)}{2\ i} y_{n-2} - ... + (-1)^n y_0 \quad (6.20)$$

The backward differences of various orders can be computed and placed in a diagonal difference table. The backward differences at a point are then found along the backward diagonal through the point. The following table shows the backward differences entries.

Diagonal difference Table of backward differences:

| $i$ | $x_i$ | $y_i$ | $\nabla y_i$ | $\nabla^2 y_i$ | $\nabla^3 y_i$ | $\nabla^4 y_i$ | $\nabla^5 y_i$ |
|---|---|---|---|---|---|---|---|
| 0 | $x_0$ | $y_0$ | | | | | |
| | | | $\nabla y_1$ | | | | |
| 1 | $x_1$ | $y_1$ | | $\nabla^2 y_2$ | | | |
| | | | $\nabla y_2$ | | $\nabla^3 y_3$ | | |
| 2 | $x_2$ | $y_2$ | | $\nabla^2 y_3$ | | $\nabla^4 y_4$ | |
| | | | $\nabla y_3$ | | $\nabla^3 y_4$ | | |
| 3 | $x_3$ | $y_3$ | | $\nabla^2 y_4$ | | $\nabla^4 y_5$ | |
| | | | $\nabla y_4$ | | $\nabla^3 y_5$ | | |
| 4 | $x_4$ | $y_4$ | | $\nabla^2 y_5$ | | | |
| | | | $\nabla y_5$ | | | | |
| 5 | $x_5$ | $y_5$ | | | | | |

NOTES

The entries along a column in the table are computed (as discussed in previous example) as the differences of the entries in the previous column and are placed in between. We notice that the backward differences of various orders with respect to $y_i$ are along the backward diagonal through it. Thus, $\nabla y_5, \nabla^2 y_5, \nabla^3 y_5, \nabla^4 y_5$ and $\nabla^5 y_5$ are along the lowest backward diagonal through $y_5$.

We may note that the data entries of the backward difference table in any column are the same as those of the forward difference table, but the differences are for different reference points.

Specifically, if we compare the columns of first order differences we can see that,

$$\Delta y_0 = \nabla y_1,\ \Delta y_1 = \nabla y_2, ...,\ \Delta y_{n-1} = \nabla y_n$$

Hence, $\Delta y_i = \nabla y_{i+1}$, for $i = 0, 1, 2, ..., n-1$

Similarly, $\Delta^2 y_0 = \nabla^2 y_2,\ \Delta^2 y_1 = \nabla^2 y_3, ...,\ \Delta^2 y_{n-2} = \nabla^2 y_n$

Thus, $\Delta^2 y_i = \nabla^2 y_{i+2}$, for $i = 1, 2, ..., n-2$

In general, $\Delta^k y_i = \nabla^k y_{i+k}$.

Conversely, $\nabla^k y_i = \Delta^k y_{i-k}$.

**Example 13:** Given the following table of values of $y = f(x)$:

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 1 | 3 | 5 | 7 | 9 |
| $y$ | 8 | 12 | 21 | 36 | 62 |

Find the values of $\nabla y_{(7)}, \nabla^2 y_{(9)}, \nabla^3 y_{(9)}$.

**Solution:** We form the diagonal difference table,

| $x_i$ | $y_i$ | $\nabla y_i$ | $\nabla^2 y_i$ | $\nabla^3 y_i$ | $\nabla^4 y_i$ |
|---|---|---|---|---|---|
| 1 | 8 | | | | |
| | | 4 | | | |
| 3 | 12 | | 5 | | |
| | | 9 | | 1 | |
| 5 | 21 | | 6 | | 4 |
| | | 15 | | 5 | |
| 7 | 36 | | 11 | | |
| | | 26 | | | |
| 9 | 62 | | | | |

From the table, we can easily find $\nabla y_{(7)} = 15$, $\nabla^2 y_{(9)} = 11$, $\nabla^3 y_{(9)} = 5$.

### Symbolic Operators

We consider the finite differences of an equally spaced tabular data for developing numerical methods. Let a function $y = f(x)$ has a set of values $y_0, y_1, y_2,...$, corresponding to points $x_0, x_1, x_2,...$, where $x_1 = x_0 + h$, $x_2 = x_0 + 2h$,...., are equally spaced with spacing $h$. We define different types of finite differences such as forward differences, backward differences and central differences, and express them in terms of operators.

The forward difference of a function $f(x)$ is defined by the operator $\Delta$, called the forward difference operator given by,

$$\Delta f(x) = f(x+h) - f(x) \tag{6.21}$$

At a tabulated point $x_i$, we have

$$\Delta f(x_i) = f(x_i + h) - f(x_i) \tag{6.22}$$

We also denote $\Delta f(x_i)$ by $\Delta y_i$, given by

$$\Delta y_i = y_{i+1} - y_i, \quad \text{for} \quad i = 0, 1, 2, ... \tag{6.23}$$

We also define an operator $E$, called the shift operator which is given by,

$$E\,f(x) = f(x + h) \tag{6.24}$$

$$\therefore \qquad \Delta f(x) = Ef(x) - f(x)$$

Thus, $\Delta = E - 1$ is an operator relation. (6.25)

While Equation (6.21) defines the first order forward difference, we can define second order forward difference by,

$$\Delta^2 y_i = \Delta(\Delta y_i) = \Delta(y_{i+1} - y_i)$$

$$\therefore \qquad \Delta^2 y_i = \Delta y_{i+1} - \Delta y_i \tag{6.26}$$

### Shift Operator

The shift operator is denoted by $E$ and is defined by $E\,f(x) = f(x + h)$. Thus,

$$Ey_k = y_{k+1}$$

Higher order shift operators can be defined by, $E^2 f(x) = Ef(x + h) = f(x + 2h)$.

$$E^2 y_k = E(Ey_k) = E(y_{k+1}) = y_{k+2}$$

In general, $$E^m f(x) = f(x + mh)$$

$$E^m yk = y_{k+m}$$

NOTES

**Relation between Forward Difference Operator and Shift Operator**

From the definition of forward difference operator, we have

$$\begin{aligned}\Delta y(x) &= y(x+h) - y(x)\\ &= Ey(x) - y(x)\\ &= (E-1)y(x)\end{aligned}$$

This leads to the operator relation,

$$\Delta = E - 1$$

or, $$E = 1 + \Delta \qquad (6.27)$$

Similarly, for the second order forward difference, we have

$$\begin{aligned}\Delta^2 y(x) &= \Delta y(x+h) - \Delta y(x)\\ &= y(x+2h) - 2y(x+h) + y(x)\\ &= E^2 y(x) - 2Ey(x) + y(x)\\ &= (E^2 - 2E + 1)y(x)\end{aligned}$$

This gives the operator relation, $\Delta^2 = (E-1)^2$.

Finally, we have $\Delta^m = (E-1)^m$, for $m = 1, 2, ...$ (6.28)

**Relation between the Backward Difference Operator with Shift Operator**

From the definition of backward difference operator, we have

$$\begin{aligned}\nabla f(x) &= f(x) - f(x-h)\\ &= f(x) - E^{-1} f(x) = (1 - E^{-1})\, f(x)\end{aligned}$$

This leads to the operator relation, $\nabla \equiv 1 - E^{-1}$ (6.29)

Similarly, the second order backward difference is defined by,

$$\begin{aligned}\nabla^2 f(x) &= \nabla f(x) - \nabla f(x-h)\\ &= f(x) - f(x-h) - f(x-h) + f(x-2h)\\ &= f(x) - 2f(x-h) + f(x-2h)\\ &= f(x) - E^{-1} f(x) + E^{-2} f(x)\\ &= (1 - E^{-1} + E^{-2})\, f(x)\\ &= (1 - E^{-1})^2\, f(x)\end{aligned}$$

This gives the operator relation, $\nabla^2 \equiv (1 - E^{-1})^2$ and in general,

$$\nabla^m \equiv (1 - E^{-1})^m \qquad (6.30)$$

### Relations between the Operators E, D and Δ

We have by Taylor's theorem,

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!} f''(x) + ...$$

Thus, $Ef(x) = f(x) + hDf(x) + \frac{h^2 D^2}{2!} f(x) + ...$, where $D = \frac{d}{dx}$

Or, $$(1+\Delta) f(x) = \left(1 + hD + \frac{h^2 D^2}{2!} + ...\right) f(x)$$
$$= e^{hD} f(x)$$

Thus, $$e^{hD} = 1 + \Delta = E \qquad (6.31)$$

Also, $$hD = \log \quad (1+\Delta)$$

Or, $$hD = \Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + ...$$

∴ $$D = \frac{1}{h} \left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + ...\right)$$

### Central Difference Operator

The central difference operator denoted by δ is defined by,

$$\delta y(x) = y\left(x + \frac{h}{2}\right) - y\left(x - \frac{h}{2}\right)$$

Thus,

$$\delta y(x) = (E^{1/2} - E^{-1/2}) y(x)$$

Giving the operator relation, $\delta = E^{1/2} - E^{-1/2}$ or $\delta E^{1/2} = E - 1$

Also,

$$\delta y_n = (E^{1/2} - E^{-1/2}) y(x_n) = E^{1/2} y_n - E^{-1/2} y_n$$

i.e., $$\delta y_n = y_{n+1/2} - y_{n-1/2}$$

Further,

$$\delta^2 y_n = \delta\,(\delta\, y_n) = \delta\, y_{n+1/2} - \delta\, y_{n-1/2}$$
$$= \left(E^{1/2} - E^{-1/2}\right) \left(y_{n+1/2}\right) - \left(E^{1/2} - E^{-1/2}\right) \left(y_{n-1/2}\right)$$
$$= E^{1/2}\left(y_{n+1/2} - y_{n-1/2}\right) - E^{-1/2}\left(y_{n+1/2} - y_{n-1/2}\right)$$
$$= y_{n+1} - y_n - y_n + y_{n-1}$$
$$= y_{n+1} - 2y_n + y_{n-1} = (E^{1/2} - E^{-1/2})^2 y_n \quad = \left[\Delta^2 y_{n-1} \equiv \nabla^2 y_{n+1}\right]$$
$$= (E + E^{-1} - 2) y_n$$

$$\therefore \quad \delta^2 \equiv E + E^{-1} - 2 \qquad (6.32)$$

NOTES

Even though the central difference operator uses fractional arguments, still it is widely used. This is related to the averaging operator and is defined by,

$$\mu = \frac{1}{2}(E^{1/2} + E^{-1/2}) \qquad (6.33)$$

Squaring, $\mu^2 = \frac{1}{4}(E + 2 + E^{-1}) = \frac{1}{4}(\delta^2 + 2 + 2) = 1 + \frac{1}{4}\delta^2$

$$\therefore \mu^2 = 1 + \frac{1}{4}\delta^2 \qquad (6.34)$$

It may be noted that, $\delta\, y_{1/2} = y_1 - y_0 = \nabla y_1$

Also, $\delta\, E^{1/2} y_1 = \delta\, y_{\frac{1}{2}+1} = y_2 - y_1 = \Delta y_1$

$$\therefore \ \delta E^{1/2} = \Delta = E - 1 \qquad (6.35)$$

Further,

$$\delta^3 y_n = \delta(\delta^2 y_n) = \delta\left(\delta y_{n+\frac{1}{2}} - \delta\, y_{n-\frac{1}{2}}\right)$$

$$= \delta^2 y_{n+\frac{1}{2}} - \delta^2 y_{n-\frac{1}{2}} = \delta(y_{n+1} - 2y_n + y_{n-1})$$

**Example 14:** Prove the following operator relations:

($i$) $\Delta \equiv \nabla E$ ($ii$) $(1+\Delta)\ (1-\nabla) = 1$

**Solution:**

($i$) Since, $\Delta f(x) = f(x+h) - f(x) = Ef(x) - f(x)$, $\Delta \equiv E - 1$ (1)

and since $\nabla f(x) = f(x) - f(x-h) = (1 - E^{-1}) f(x)$, $\nabla \equiv 1 - E^{-1}$ (2)

Thus, $\nabla \equiv \frac{E-1}{E}$ or $\nabla E \equiv E - 1 \equiv \Delta$

Hence proved.

($ii$) From Equation (1), we have $E \equiv \Delta + 1$ (3)

and from Equation (2) we get $E^{-1} \equiv 1 - \nabla$ (4)

Combining Equations (3) and (4), we get $(1+\Delta)(1-\nabla) \equiv 1$.

**Example 15:** If $f_i$ is the value of $f(x)$ at $x_i$ where $x_i = x_0 + ih$, for $i = 1,2,...,$ prove that,

$$f_i = E^i f_o = \sum_{j=0}^{i} \binom{i}{j} \Delta^i f_0$$

NOTES

**Solution:** We can write $Ef(x) = f(x+h)$

Using Taylor series expansion, we have

$$Ef(x) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + ...$$

$$= f(x) + hDf(x) + \frac{h^2}{2!}D^2 f(x) + ..., \quad \text{where} \quad D = \frac{d}{dx}$$

$$\therefore \quad (1+\Delta)f(x) = \left(1 + hD + \frac{h^2D^2}{2!} + ...\right)f(x), \quad \text{since} \quad E = 1+\Delta$$

$$= e^{hD}.f(x)$$

$$\therefore \quad 1+\Delta = e^{hD}$$

Hence, $e^{ihD} = (1+\Delta)^i$

Now, $f_i = f(x_i) = f(x_0 + ih) = E^i f(x_0)$

$$\therefore \quad f_i = (1+\Delta)^i f(x_0), \quad \text{since} \quad E \equiv 1+\Delta$$

$$f_i = \sum_{j=0}^{i} \binom{i}{j} \Delta^i f_0, \text{ using binomial expansion.}$$

Hence proved.

**Example 16:** Compute the following differences:

$(i)$ $\Delta^n e^x$ $(ii)$ $\Delta^n x^n$

**Solution:**

$(i)$ We have, $\Delta e^x = e^{x+h} - e^x = e^x(e^h - 1)$

Again, $\Delta^2 e^x = \Delta(\Delta e^x) = (e^h - 1)\Delta e^x = (e^h - 1)^2 e^x$

Thus by induction, $\Delta^n e^x = (e^h - 1)^n e^x$.

$(ii)$ We have,

$$\Delta(x^n) = (x+h)^n - x^n$$

$$= n\,h\,x^{n-1} + \frac{n(n-1)}{2!}h^2 x^{n-2} + .... + h^n$$

Thus, $\Delta(x^n)$ is a polynomial of degree $(n-1)$

Also, $\Delta(h^n) = 0$. Hence, we can say that $\Delta^2(x^n)$ is a polynomial of degree $(n-2)$ with the leading term $n(n-1)h^2 x^{n-2}$.

NOTES

Proceeding $n$ times, we get

$$\Delta^n(x^n) = n(n-1)\ldots 1h^n = n!h^n$$

**Example 17:** Prove that,

$(i)$ $\Delta\left\{\frac{f(x)}{g(x)}\right\} = \frac{g(x)\,\Delta f(x) - f(x)\,\Delta g(x)}{g(x)\,g(x+h)}$

$(ii)$ $\Delta\{\log f(x)\} = \log\left\{1 + \frac{\Delta f(x)}{f(x)}\right\}$

**Solution:**

$(i)$ We have,

$$\Delta\left\{\frac{f(x)}{g(x)}\right\} = \frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)}$$

$$= \frac{f(x+h)\,g(x) - f(x)g(x+h)}{g(x+h)\,g(x)}$$

$$= \frac{f(x+h)\,g(x) - f(x)\,g(x) + f(x)\,g(x) - f(x)\,g(x+h)}{g(x+h)\,g(x)}$$

$$= \frac{g(x)\{f(x+h) - f(x)\} - f(x)\{g(x+h) - g(x)\}}{g(x)\,g(x+h)}$$

$$= \frac{g(x)\,\Delta f(x) - f(x)\Delta g(x)}{g(x)\,g(x+h)}$$

$(ii)$ We have,

$$\Delta\{\log f(x)\} = \log\{f(x+h)\} - \log\{f(x)\}$$

$$= \log\frac{f(x+h)}{f(x)} = \log\left\{\frac{f(x+h) - f(x) + f(x)}{f(x)}\right\}$$

$$= \log\left\{\frac{\Delta f(x)}{f(x)} + 1\right\}$$

### Differences of a Polynomial

We now look at the differences of various orders of a polynomial of degree $n$, given by

$$y = f(x) = a_n x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_1 x + a_0$$

The first order forward difference is defined by,

$\Delta f(x) = f(x+h) - f(x)$ and is given by,

$$\Delta y = a_n\{(x+h)^n - x^n\} + a_{n-1}\{(x+h)^{n-1} - x^{n-1}\} + \ldots + a_1(x+h-x)$$

$$= a_n\{n\,h\,x^{n-1} + \frac{n(n-1)}{2\,!}h^2 x^{n-2} + \ldots\} + a_{n-1}\{(n-1)h\,x^{n-2} + \ldots\}$$

$$= b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \ldots + b_1 x + b_0$$

NOTES

where the coefficients of various powers of $x$ are collected separately.

Thus, the first order difference of a polynomial of degree $n$ is a polynomial of degree $n-1$, with $b_{n-1} = a_n.nh$

Proceeding as above, we can state that the second order forward difference of a polynomial of degree $n$ is a polynomial of degree $n-2$, with coefficient of $x^{n-2}$ as $n(n-1)h^2a_0$.

Continuing successively, we finally get $\Delta^n y = a_0 n! h^n$, a constant.

We can conclude that for polynomial of degree $n$, all other differences having order higher than $n$ are zero.

It may be noted that the converse of the above result is partially true and suggests that if the tabulated values of a function are found to be such that the differences of the $k$th order are approximately constant, then the highest degree of the interpolating polynomial that should be used is $k$. Since the tabulated data may have round-off errors, the acutal function may not be a polynomial.

**Example 18:** Compute the horizontal difference table for the following data and hence, write down the values of $\nabla f(4)$, $\nabla^2 f(3)$ and $\nabla^3 f(5)$.

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $f(x)$ | 3 | 18 | 83 | 258 | 627 |

**Solution:** The horizontal difference table for the given data is as follows:

| $x$ | $f(x)$ | $\nabla f(x)$ | $\nabla^2 f(x)$ | $\nabla^3 f(x)$ | $\nabla^4 f(x)$ |
|---|---|---|---|---|---|
| 1 | 3 | – | – | – | – |
| 2 | 18 | 15 | – | – | – |
| 3 | 83 | 65 | 50 | – | – |
| 4 | 258 | 175 | 110 | 60 | – |
| 5 | 627 | 369 | 194 | 84 | 24 |

From the table we read the required values and get the following result:

$$\nabla f(4) = 175,\ \nabla^2 f(3) = 50,\ \nabla^3 f(5) = 84$$

**Example 19:** Form the difference table of $f(x)$ on the basis of the following table and show that the third differences are constant. Hence, conclude about the degree of the interpolating polynomial.

| $x$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $f(x)$ | 5 | 6 | 13 | 32 | 69 |

NOTES

**Solution:** The difference table is given below

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ |
|---|---|---|---|---|
| 0 | 5 | | | |
| | | 1 | | |
| 1 | 6 | | 6 | |
| | | 7 | | 6 |
| 2 | 13 | | 12 | |
| | | 19 | | 6 |
| 3 | 32 | | 18 | |
| | | 37 | | |
| 4 | 69 | | | |

It is clear from the above table that the third differences are constant and hence, the degree of the interpolating polynomial is three.

**Newton's Forward Difference Interpolation Formula**

Newton's forward difference interpolation formula is a polynomial of degree less than or equal to $n$. This is used to find the value of the tabulated function at a non-tabular point. Consider a function $y = f(x)$ whose values $y_0, y_1, ..., y_n$ at a set of equidistant points $x_0, x_1, ..., x_n$ are known.

Let $\varphi(x)$ be the interpolating polynomial, such that

$$\varphi(x_i) = f(x_i) = y_i$$
$$x_i = x_0 + ih, \quad \text{for} \quad i = 0, 1, 2, ..., n \qquad (6.36)$$

We assume the polynomial $\varphi(x)$ to be of the form,

$$\varphi(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2) + \\ ... + a_n(x - x_0)(x - x_1)...(x - x_{n-1}) \qquad (6.37)$$

The coefficients $a_i$'s in Equation (6.37) are determined by satisfying the conditions in Equation (6.36) successively for $i = 0, 1, 2, ..., n$.

Thus, we get

$$y_0 = \varphi(x_0) = a_0, \quad \text{gives} \quad a_0 = y_0$$

$$y_1 = \varphi(x_1) = a_0 + a_1(x_1 - x_0), \quad \text{gives} \quad a_1 = \frac{y_1 - y_0}{h}$$

$$\therefore \quad a_1 = \frac{\Delta y_0}{h}$$

$$y_2 = \varphi(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$$

NOTES

or,

$$y_2 = y_0 + \frac{\Delta y_0}{h} + a_2(2h) \quad h$$

$$a_2 = \frac{y_2 - 2y_1 + y_0}{2h^2} = \frac{\Delta^2 y_0}{2\ !\ \ h^2}$$

Proceeding further, we get successively,

$$a_3 = \frac{\Delta^3\ y_0}{3\,!\,h^3}, \ldots, a_n = \frac{\Delta^n y_0}{n\,!\,h^n}$$

Using these values of the coefficients, we get Newton's forward difference interpolation in the form,

$$\varphi(x) = y_0 + \frac{x - x_0}{h}\Delta\ y_0 + \frac{(x - x_0)(x - x_1)}{2\ !\ \ h^2}\Delta^2 y_0 + \frac{(x - x_0)}{h}\frac{(x - x_1)}{h}\frac{(x - x_2)}{h}\frac{\Delta^3 y_0}{3!} + \ldots$$

$$+ \ldots + \frac{(x - x_0)}{h}\frac{(x - x_1)}{h}\ldots\frac{(x - x_{n-1})}{h}\frac{\Delta^n y_0}{n!}$$

This formula can be expressed in a more convenient form by taking $u = \frac{x - x_0}{h}$ as shown here.

We have,

$$\frac{x - x_1}{h} = \frac{x - (x_0 + h)}{h} = \frac{x - x_0}{h} - 1 = u - 1$$

$$\frac{x - x_2}{h} = \frac{x - (x_0 + 2h)}{h} = \frac{x - x_0}{h} - 2 = u - 2$$

$$\frac{x - x_{n-1}}{h} = \frac{x - \{x_0 + (n-1)h\}}{h} = \frac{x - x_0}{h} - (n-1) = u - n + 1$$

Thus, the interpolating polynomial reduces to:

$$\varphi(u) = y_0 + u \quad \Delta y_0 + \frac{u(u-1)}{2\ !}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3\ !}\Delta^3 y_0$$

$$+ \ldots + \frac{u(u-1)(u-2)\ldots(u-n+1)}{n\ !}\Delta^n y_0 \qquad (6.38)$$

This formula is generally used for interpolating near the beginning of the table. For a given $x$, we choose a tabulated point as $x_0$ for which the following condition is satisfied.

For better results, we should have

$$|u| = \left|\frac{x - x_0}{h}\right| \le 0.5$$

NOTES

The degree of the interpolating polynomial to be used is less than or equal to $n$ and is determined by the order of the differences when they are nearly same so that the differences of higher orders are irregular due to the propagated round-off error in the data.

**Newton's Backward Difference Interpolation Formula**

Newton's forward difference interpolation formula cannot be used for interpolating at a point near the end of a table, since we do not have the required forward differences for interpolating at such points. However, we can use a separate formula known as Newton's backward difference interpolation formula. Let a table of values $\{x_i, y_i\}$, for $i = 0, 1, 2, ..., n$ for equally spaced values of $x_i$ be given. Thus, $x_i = x_0 + ih$, $y_i = f(x_i)$, for $i = 0, 1, 2, ..., n$ are known.

We construct an interpolating polynomial of degree $n$ of the form,

$$y \approx \varphi(x) = b_0 + b_1(x - x_n) + b_2(x - x_n)(x - x_{n-1}) + ... + b_n(x - x_n)(x - x_{n-1})...(x - x_1) \quad (6.39)$$

We have to determine the coefficients $b_0, b_1, ..., b_n$ by satisfying the relations,

$$\varphi(x_i) = y_i, \quad \text{for} \quad i = n, n-1, n-2, ..., 1, 0 \quad (6.40)$$

Thus, $\varphi(x_n) = y_n$, gives $b_0 = y_n$ (6.41)

Similarly, $\varphi(x_{n-1}) = y_{n-1}$, gives $y_{n-1} = b_0 + b_1(x_{n-1} - x_n)$

Or,

$$b_1 = \frac{y_n - y_{n-1}}{h} = \frac{\nabla y_n}{h} \quad (6.42)$$

Again

$$\varphi(x_{n-2}) = y_{n-2}, \quad \text{gives} \quad y_{n-2} = b_0 + b_1(x_{n-2} - x_n) + b_2(x_{n-2} - x_n)(x_{n-1} - x_n)$$

Or,

$$y_{n-2} = y_n + \frac{y_n - y_{n-1}}{h}(-2h) + b_2(-2h)(-h)$$

$$\therefore \quad b_2 = \frac{y_n - 2y_{n-1} + y_{n-2}}{2h^2} = \frac{\nabla^2 y_n}{2!\,h^2} \quad (6.43)$$

By induction or by proceeding as mentioned earlier, we have

$$b_3 = \frac{\nabla^3 y_n}{3\,!\;h^3},\ b_4 = \frac{\nabla^4 y_n}{4\,!\;h^4},\ ...,\ b_n = \frac{\nabla^n y_n}{n\,!\;h^n} \quad (6.44)$$

Substituting the expressions for $b_i$ in Equation (6.39), we get

$$\varphi(x) = y_n + \frac{\nabla y_n}{h}(x - x_n) + \frac{\nabla^2 y_n}{2\,!\;h^2}(x - x_n)\quad(x - x_{n-1}) + ... + \frac{\nabla^n y_n}{n\,!\;h^n}(x - x_n)(x - x_{n-1})...(x - x_1) \quad (6.45)$$

This formula is known as Newton's backward difference interpolation formula. It uses the backward differences along the backward diagonal in the difference table.

Introducing a new variable $v = \dfrac{x - x_n}{h}$,

we have, $\dfrac{x - x_{n-1}}{h} = \dfrac{x - (x_n - h)}{h} = v + 1$.

Similarly, $\dfrac{x - x_{n-2}}{h} = v + 2, ..., \dfrac{x - x_1}{h} = v + \overline{n-1}$.

Thus, the interpolating polynomial in Equation (6.45) may be rewritten as,

$$\varphi(x) = y_n + v\nabla y_n + \frac{v(v+1)}{2\,!}\nabla^2 y_n + \frac{v(v+1)(v+2)}{3\,!}\nabla^3 y_n + ... + \frac{v(v+1)(v+2)...(v+n-1)}{n\,!}\nabla^n y_n \quad (6.46)$$

This formula is generally used for interpolation at a point near the end of a table.

The error in the given interpolation formula may be written as,

$$E(x) = f(x) - \varphi(x)$$

$$= \frac{(x - x_n)(x - x_{n-1})...(x - x_1)(x - x_0)\,f^{(n+1)}(\xi)}{(n+1)\,!}, \quad \text{where} \quad x_0 < \xi < x_n$$

$$= v(v+1)(v+2)...(v+n)\frac{y^{(n+1)}(\xi)}{(n+1)\,!}h^{n+1}$$

### Extrapolation

The interpolating polynomials are usually used for finding values of the tabulated function $y = f(x)$ for a value of $x$ within the table. But, they can also be used in some cases for finding values of $f(x)$ for values of $x$ near to the end points $x_0$ or $x_n$ outside the interval $[x_0, x_n]$. This process of finding values of $f(x)$ at points beyond the interval is termed as extrapolation. We can use Newton's forward difference interpolation for points near the beginning value $x_0$. Similarly, for points near the end value $x_n$, we use Newton's backward difference interpolation formula.

**Example 20:** With the help of appropriate interpolation formula, find from the following data the weight of a baby at the age of one year and of ten years:

| Age = $x$ | 3 | 5 | 7 | 9 |
|---|---|---|---|---|
| Weight = $y$ (kg) | 5 | 8 | 12 | 17 |

**Solution:** Since the values of $x$ are equidistant, we form the finite difference table for using Newton's forward difference interpolation formula to compute weight of the baby at the age of required years.

NOTES

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ |
|---|---|---|---|
| 3 | 5 | | |
| | | 3 | |
| 5 | 8 | | 1 |
| | | 4 | |
| 7 | 12 | | 1 |
| | | 5 | |
| 9 | 17 | | |

Taking $x = 2$, $u = \dfrac{x - x_0}{h} = -0.5.$

Newton's forward difference interpolation gives,

$$y \text{ at } x = 1,\ y(1) = 5 - 0.5 \times 3 + \frac{(-0.5)(-1.5)}{2} \times 1$$

$$= 5 - 1.5 + 0.38 = 3.88 \simeq 3.9 \text{ kg.}$$

Similarly, for computing weight of the baby at the age of ten years, we use Newton's backward difference interpolation given by,

$$v = \frac{x - x_n}{h} = \frac{10 - 9}{2} = 0.5$$

$$y \text{ at } x = 10,\ y(10) = 17 + 0.5 \times 5 + \frac{0.5 \times 1.5}{2} \times 1$$

$$= 17 + 2.5 + 0.38 \simeq 19.88$$

### Inverse Interpolation

The problem of inverse interpolation in a table of values of $y = f(x)$ is to find the value of $x$ for a given $y$. We know that the inverse function $x = g(y)$ exists and is unique, if $y = f(x)$ is a single valued function of $x$ and $\dfrac{dy}{dx}$ exists and does not vanish in the neighbourhood of the point where inverse interpolation is desired.

When the values of $x$ are unequally spaced, we can apply Lagrange's interpolation or iterative linear interpolation simply by interchanging the roles of $x$ and $y$. Thus Lagrange's formula for inverse interpolation can be written as,

$$x = \sum_{i=0}^{n} l_i(y) x_i$$

Where
$$l_i(y) = \prod_{\substack{j=0 \\ j \neq i}}^{n} [(y - y_j)/(y_i - y_j)]$$

When $x$ values are equally spaced, we can apply the method of successive approximation as described below.

Consider Newton's formula for forward difference interpolation given by,

$$y = y_0 + u\,\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + \ldots$$

Retaining only two terms on the RHS, we can write the first approximation,

$$u^{(1)} = \frac{1}{\Delta y_0}(y - y_0)$$

The second approximation can be written as,

$$u^{(2)} = \frac{1}{\Delta y_0}\left[(y - y_0) - \frac{u^{(1)}(u^{(1)} - 1)}{2}\Delta^2 y_0\right]$$

on replacing $u$ by $u^{(1)}$ in the coefficient of $\Delta^2 y_0$.

Similarly, the third approximation can be written as,

$$u^{(3)} = \frac{1}{\Delta y_o}\left[y - y_0 - \frac{u^{(2)}(u^{(2)} - 1)}{2}\Delta^2 y_0 - \frac{u^{(2)}(u^{(2)} - 1)(u^{(2)} - 2)}{6}\Delta^3 y_0\right]$$

The process can be continued until two successive approximations have a reasonable accuracy. Then $x$ is obtained by the relation,

$$x = x_0 + uh$$

**Example 21:** Using inverse interpolation, find the value of $x$ for $y = 5$, from the given table.

| | | | |
|---|---|---|---|
| $x$ | 1 | 3 | 4 |
| $y$ | 3 | 12 | 19 |

**Solution:** Applying inverse interpolation,

$$x = \sum_{i=0}^{2} l_i(y).x_i$$

Thus, for $y = 5$, we have

$$x = \frac{(5-12)(5-19)}{(3-12)(3-19)} \times 1 + \frac{(5-3)(5-19)}{(12-3)(12-19)} \times 3 + \frac{(5-3)(5-12)}{(19-3)(19-12)} \times 4$$

$$= \frac{7 \times 14}{9 \times 16} + \frac{2 \times -14}{9 \times -7} \times 3 + \frac{2 \times -7}{16 \times 7} \times 4$$

$$= 0.6805 + 1.3333 - 0.5000$$

$$= 1.5138$$

$$= 1.514 \text{ correct upto four significant figures.}$$

**Example 22:** Given the following tabular values of cos$h$ $x$, find $x$ for which cos$h$ $x$ = 1.285.

| $x$ | 0.738 | 0.739 | 0.740 | 0.741 | 0.742 |
|---|---|---|---|---|---|
| cos$h$ $x$ | 1.2849085 | 1.2857159 | 1.2865247 | 1.2873348 | 1.2881461 |

**NOTES**

**Solution:** Since finding $x$ for an equally spaced table of cos$h$ $x$ is a problem of inverse interpolation, we employ the method of successive approximation using Newton's formula of inverse interpolation. We first form the finite difference table.

| $x$ | $f(x) = \cos h\, x$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ |
|---|---|---|---|---|
| 0.738 | 1.2849085 | | | |
| | | 8074 | | |
| 0.739 | 1.2857159 | | 14 | |
| | | 8088 | | –1 |
| 0.740 | 1.2865247 | | 13 | |
| | | 8101 | | –1 |
| 0.741 | 1.2873348 | | 12 | |
| | | 8113 | | |
| 0.742 | 1.2881461 | | | |

Using Newton's forward difference interpolation formula for the first approximation $u = \dfrac{(x - x_0)}{h}$ we get,

$$u^{(1)} = \frac{1}{\Delta f(x_0)}(y - y_0)$$

For, $\quad y = 1.285,$ we take $x_0 = 0.739.$

$\therefore$

$$u^{(1)} = \frac{1}{0.0008088} \times (1.285 - 1.2857159) = -0.8851384, \text{ then } x \approx 0.739885$$

For a second approximation,

$$u^{(2)} = u^{(1)} - \frac{1}{\Delta f(x_0)} \frac{u^{(1)}(u^{(1)} - 1)}{2} \Delta^2 f(x_0)$$

$$= -0.8851384 - \frac{1}{0.0008088 \times 2} \times (-0.8851384) \times (-1.8851384) \times 0.0000013$$

$$= -0.8851384 - 0.0013409 = -0.8864793 \Rightarrow x = 0.7398864$$

Similarly,

$$u^{(3)} = u^{(1)} - \frac{u^{(2)}(u^{(2)} - 1)}{2} \frac{\Delta^2 f_0}{\Delta f_0} - \frac{1}{6} u^{(2)}(u^{(2)} - 1)(u^{(2)} - 2) \frac{\Delta^3 f_0}{\Delta f_0}$$

$$= -0.8851384 - 0.0013430 - 0.000073600$$

$$= -0.886555 \Rightarrow x = 0.7398865$$

**Example 23:** Find the divided difference interpolation for the following table of values:

| $x$ | 4 | 7 | 9 |
|---|---|---|---|
| $f(x)$ | –43 | 83 | 327 |

**Solution:** We first form the Divided Difference (DD) table as given below.

| $x$ | $f(x)$ | 1st $DD$ | 2nd $DD$ |
|---|---|---|---|
| 4 | $-43$ | | |
| | | 42 | |
| 7 | 83 | | 16 |
| | | 122 | |
| 9 | 327 | | |

Newton's divided difference interpolation formula is,

$$f(x) \approx f(x_0) + (x - x_0)\, f(x_0, x_1) + (x - x_0)\,(x - x_1)\, f\left[x_0,\ x_1,\ x_2\right]$$

$$\therefore f(x) \approx -43 + (x-4)\,42 + (x-4)\,(x-7) \times 16$$

$$= 16x^2 - 134x + 237$$

**Example 24:** Given the following table of values of the function $y = \log_e x$, construct the Newton's forward difference interpolating polynomial. Comment on the degree of the polynomial and find $\log_e 1001$.

| $x$ | 1000 | 1010 | 1020 | 1030 | 1040 |
|---|---|---|---|---|---|
| $\log_e x$ | 3.00000 | 3.00432 | 3.00860 | 3.01284 | 3.01703 |

**Solution:** We form the difference table as given below:

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ |
|---|---|---|---|
| 1000 | 3.00000 | | |
| | | 432 | |
| 1010 | 3.00432 | | $-4$ |
| | | 428 | |
| 1020 | 3.00860 | | $-4$ |
| | | 424 | |
| 1030 | 3.01284 | | $-5$ |
| | | 419 | |
| 1040 | 3.01703 | | |

We observe that, the differences of second order are nearly constant. Thus, the degree of the interpolating polynomial is 2 and is given by,

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{2}\Delta^2 y_0, \text{ where } u = \frac{x - x_0}{h}$$

For $x = 1001$, we take $x_0 = 1000$.

NOTES

$$\therefore u = \frac{1001 - 1000}{10} = 0.1$$

$$\log_e 1001 = 3.00000 + 0.1 \times 0.00432 + \frac{0.1 \times 0.9}{2} \times (-0.00004)$$

$$= 3.000430 \approx 3.00043$$

**Example 25:** Determine the interpolating polynomial for the following data table using both forward and backward difference interpolating formulae. Comment on the result.

| $x$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $f(x)$ | 1.0 | 8.5 | 36.0 | 95.5 | 199.0 |

**Solution:** Since the data points are equally spaced, we construct the Newton's forward difference interpolating polynomial for which we first form the finite difference table as given below:

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ |
|---|---|---|---|---|
| 0 | 1.0 | | | |
| | | 7.5 | | |
| 1.0 | 8.5 | | 20.0 | |
| | | 27.5 | | 12.0 |
| 2.0 | 36.0 | | 32.0 | |
| | | 59.5 | | 12.0 |
| 3.0 | 95.5 | | 44.0 | |
| | | 103.5 | | |
| 4.0 | 199.0 | | | |

Since the differences of order 3 are constant, we construct the third degree Newton's forward difference interpolating polynomial given by,

$$f(x) \cong 1.0 + x \times 0.75 + \frac{x(x-1)}{2} \times 20 + \frac{x(x-1)(x-2)}{6} \times 12$$

*Since* $x_0 = 0, \quad h = 1.0$

$$\therefore \qquad u = \frac{x - x_0}{h} = x$$

*i.e.,* $f(x) = 1.0 + 1.5x + 4x^2 + 2x^3$, *on simplification.*

Taking $x_n = 4$, we also construct the backward difference interpolating polynomial given by,

NOTES

$$f(x) = 199 + (x-4)\times 103.5 + \frac{(x-4)\ \ (x-3)}{2}\times 44$$
$$+\frac{(x-4)\ \ (x-3)\ \ (x-3)}{6}\times 12$$
$$= 1.0 + 1.5x + 4x^2 + 2x^3,\text{ on simplification.}$$

This is the same as the forward difference interpolating polynomial, because the difference of third order is constant.

**Example 26:** Use Newton's divided difference interpolation to evaluate $f(18)$ and $f(12)$ for the following data:

| $x$ | 4 | 5 | 7 | 10 | 11 | 13 |
|---|---|---|---|---|---|---|
| $f(x)$ | 48 | 100 | 294 | 900 | 1210 | 2028 |

**Solution:** We first form the divided difference table as given below.

| $x$ | $f(x)$ | 1st $DD$ | 2nd $DD$ | 3rd $DD$ |
|---|---|---|---|---|
| 4 | 48 | | | |
| | | 52 | | |
| 5 | 100 | | 15 | |
| | | 97 | | 1 |
| 7 | 294 | | 21 | |
| | | 202 | | 1 |
| 10 | 900 | | 27 | |
| | | 310 | | 1 |
| 11 | 1210 | | 33 | |
| | | 409 | | |
| 13 | 2028 | | | |

Since 3rd order divided differences are same, higher order divided differences vanish. We have the Newton's divided difference interpolation given by,

$$f(x) \approx f_0 + (x-x_0)f[x, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2]$$
$$+ (x-x_0)(x-x_1)(x-x_2)f[x_0, x_1, x_2, x_3]$$

For $x = 8$, we take $x_0 = 4$,

$$f(8) = 48 + (8-4)52 + (8-4)(8-5)\times 15 + (8-4)(8-5)(8-7)\times 1$$
$$= 48 + 208 + 180 + 12 = 448$$

For $x = 12$, we take $x_0 = 13$,

$$f(12) = 2028 + (12-13)\times 409 + (12-13)\ (12-11)\times 33$$
$$+ (12-13)\ (12-11)\ (12-10)\times 1$$
$$\therefore f(12) = 2028 - 409 - 33 - 2 = 1584$$

NOTES

**Example 27:** Using inverse interpolation, find the zero of $f(x)$ given by the following tabular values.

| $x$ | 0.3 | 0.4 | 0.6 | 0.7 |
|---|---|---|---|---|
| $y = f(x)$ | 0.14 | 0.06 | $-0.04$ | $-0.06$ |

**Solution:** Using Lagrange's form of inverse interpolation, we calculate the formula using $y = 0.14, 0.06, -0.04$ and $-0.06$, as given below:

$$P_3(y) = \frac{(y-0.06)(y+0.04)(y+0.06)}{(0.14-0.06)(0.14+0.04)(0.14+0.06)} \times 0.3 + \frac{(y-0.14)(y+0.04)(y+0.06)}{(0.06-0.14)(0.06+0.04)(0.06+0.06)} \times 0.4 + \frac{(y-0.14)(y-0.06)(y+0.06)}{(0.04-0.14)(0.04-0.06)(-0.04+0.06)} \times 0.6 + \frac{(y-0.14)(y-0.06)(y+0.04)}{(-0.06-0.14)(-0.06-0.06)(-0.06+0.04)} \times 0.7$$

Thus, $$P_3(0) = \frac{-0.06 \times 0.04 \times 0.06 \times 0.3}{0.08 \times 0.18 \times 0.20} + \frac{0.14 \times 0.04 \times 0.06 \times 0.4}{0.08 \times 0.1 \times 0.12} + \frac{0.14 \times 0.06 \times 0.06 \times 0.6}{0.18 \times 0.1 \times 0.02} - \frac{0.14 \times 0.06 \times 0.04 \times 0.7}{0.2 \times 0.12 \times 0.02}$$
$$= -0.015 + 0.14 + 0.84 - 0.49 = 0.475$$

Thus, the zero of $f(x)$ is 0.475 which is approximately equal to 0.48, since the accuracy depends on the accuracy of the data which is the significant digits.

### Hermite Interpolation

**Hermite Interpolation:** Hermite interpolation, named after Charles Hermite, is a method of interpolating data points as a polynomial function. The generated Hermite interpolating polynomial is closely related to the Newton polynomial, in that both are derived from the calculation of divided differences. However, the Hermite interpolating polynomial may also be computed without using divided differences, see Chinese remainder theorem and Hermite interpolation.

Unlike Newton interpolation, Hermite interpolation matches an unknown function both in observed value, and the observed value of its first $m$ derivatives. This means that $n(m + 1)$ values,

$$\begin{array}{llll} (x_0, y_0), & (x_1, y_1), & \ldots, & (x_{n-1}, y_{n-1}), \\ (x_0, y'_0), & (x_1, y'_1), & \ldots, & (x_{n-1}, y'_{n-1}), \\ \vdots & \vdots & & \vdots \\ (x_0, y_0^{(m)}), & (x_1, y_1^{(m)}), & \ldots, & (x_{n-1}, y_{n-1}^{(m)}) \end{array}$$

must be known, rather than just the first $n$ values required for Newton interpolation. The resulting polynomial may have degree at most $n(m + 1) – 1$, whereas the Newton polynomial has maximum degree $n – 1$. (In the general case, there is no need for $m$ to be a fixed value; that is, some points may have more known derivatives than others. In this case the resulting polynomial may have degree $N – 1$, with $N$ the number of data points.)

**Check Your Progress**

1. What do we generate in iterative linear interpolation?
2. Define interpolation.
3. How is Lagrange's interpolation useful?
4. Which interpolation will you use for equally spaced tabular values?
5. Define the shift operator.
6. What is the Newton forward difference interpolation formula used?
7. Define extrapolation.
8. Define the problem of inverse interpolation.

## 6.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In this method, we successively generate interpolating polynomials of any degree by iteratively using linear interpolating functions.
2. It can be stated explicitly as 'given a set of $(n + 1)$ values $y_0, y_1, y_2, ..., y_n$ for $x = x_0, x_1, x_2, ..., x_n$ respectively. The problem of interpolation is to compute the value of the function $y = f(x)$ for some non-tabular value of $x$.'
3. Lagrange's interpolation is useful for unequally spaced tabulated values.
4. For interpolation of an unknown function when the tabular values of the argument $x$ are equally spaced, we have two important interpolation formulae, viz.,
   (a) Newton's forward difference interpolation formula
   (b) Newton's backward difference interpolation formula
5. The shift operator is denoted by $E$ and is defined by $Ef(x) = f(x + h)$.
6. The Newton's forward difference interpolation formula is a polynomial of degree less than or equal to $n$.
7. The interpolating polynomials are usually used for finding values of the tabulated function $y = f(x)$ for a value of $x$ within the table. But they can also be used in some cases for finding values of $f(x)$ for values of $x$ near to the

NOTES

end points $x_0$ or $x_n$ outside the interval $[x_0, x_n]$. This process of finding values of $f(x)$ at points beyond the interval is termed as extrapolation.

8. The problem of inverse interpolation in a table of values of $y = f(x)$ is to find the value of $x$ for a given $y$.

## 6.4 SUMMARY

- The problem of interpolation is very fundamental problem in numerical analysis.
- In numerical analysis, interpolation means computing the value of a function $f(x)$ in between values of $x$ in a table of values.
- Lagrange's interpolation is useful for unequally spaced tabulated values.
- For interpolation of an unknown function when the tabular values of the argument $x$ are equally spaced, we have two important interpolation formulae, viz., Newton's forward difference interpolation formula and Newton's backward difference interpolation formula.
- The forward difference operator is defined by, $\Delta f(x) = f(x+h) - f(x)$.
- The backward difference operator is defined by, $\Delta f(x) = f(x+h) - f(x)$.
- We define different types of finite differences such as forward differences, backward differences and central differences, and express them in terms of operators.
- The shift operator is denoted by $E$ and is defined by $E f(x) = f(x + h)$.
- The first order difference of a polynomial of degree $n$ is a polynomial of degree $n$–1. For polynomial of degree $n$, all other differences having order higher than $n$ are zero.
- Newton's forward difference interpolation formula is generally used for interpolating near the beginning of the table while Newton's backward difference formula is used for interpolating at a point near the end of a table.
- In iterative linear interpolation, we successively generate interpolating polynomials, of any degree, by iteratively using linear interpolating functions.
- The process of finding values of a function at points beyond the interval is termed as extrapolation.
- The problem of inverse interpolation in a table of values of $y = f(x)$ is to find the value of $x$ for a given $y$.

## 6.5 KEY WORDS

- **Interpolation:** It means computing the value of a function $f(x)$ in between values of $x$ in a table of values.

- **Extrapolation:** The process of finding values of a function at points beyond the interval is termed as extrapolation.

## 6.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

NOTES

### Short-Answer Questions

1. What is the significance of polynomial interpolation?
2. Define the symbolic operators $E$ and $\Delta$.
3. What is the degree of the first order forward difference of a polynomial of degree $n$?
4. What is the degree of the $n$th order forward difference of a polynomial of degree $n$?
5. Write Newton's forward and backward difference formulae.
6. State an application of iterative linear interpolation.
7. What is the advantage of extrapolation?
8. State Lagrange's formula for inverse interpolation.

### Long-Answer Questions

1. Use Lagrange's interpolation formula to find the polynomials of least degree which attain the following tabular values:

(a)

| $x$ | –2 | 1 | 2 |
|---|---|---|---|
| $y$ | 25 | –8 | –15 |

(b)

| $x$ | 0 | 1 | 2 | 5 |
|---|---|---|---|---|
| $y$ | 2 | 3 | 12 | 147 |

(c)

| $x$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $y$ | –1 | –1 | 1 | 5 |

2. Form the finite difference table for the given tabular values and find the values of:

(a) $\Delta f(2)$

(b) $\Delta f^2(1)$

(c) $\Delta f^3(0)$

(d) $\Delta f^4(1)$

(e) $f(5)$

(f) $f(3)$

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $f(x)$ | 3 | 4 | 13 | 36 | 79 | 148 | 249 |

NOTES

3. How are the forward and backward differences in a table related? Prove the following:

   (a) $\Delta y_i = \nabla y_{i+1}$

   (b) $\Delta^2 y_i = \nabla^2 y_{i+2}$

   (c) $\Delta^n y_i = \nabla^n y_{i+n}$

4. Describe Newton's forward and backward difference formulae using illustrations.
5. Explain iterative linear interpolation with the help of examples.
6. Illustrate inverse interpolation procedure.

## 6.7 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

# UNIT 7 APPROXIMATION

NOTES

**Structure**



## 7.0 INTRODUCTION

Numerical error is the combined effect of two kinds of error in a calculation. The first is caused by the finite precision of computations involving floating point or integer values. The second usually called truncation error is the difference between the exact mathematical solution and the approximate solution obtained when simplifications are made to the mathematical equations to make them more amenable to calculation. The number of significant figures in a measurement, such as 2.531, is equal to the number of digits that are known with some degree of confidence (2, 5 and 3) plus the last digit (1), which is an estimate or approximation. Zeroes within a number are always significant. Zeroes that do nothing but set the decimal point are not significant. Trailing zeroes that are not needed to hold the decimal point are significant. A round-off error, also called rounding error, is the difference between the calculated approximation of a number and its exact mathematical value. Numerical analysis specifically tries to estimate this error when using approximation equations and/or algorithms, especially when using finitely many digits to represent real numbers.

In this unit, you will study about the approximation and least square approximation.

## 7.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the various types of approximations
- Evaluate errors in functions
- Define significance errors
- Understand the characteristics of numerical computation
- Analyse the least square approximation

NOTES

## 7.2 APPROXIMATION

Numerical methods are methods used for solving problems through numerical calculations providing a table of numbers and/or graphical representations or figures. Numerical methods emphasize that how the algorithms are implemented. Thus, the objective of numerical methods is to provide systematic methods for solving problems in a numerical form. Often the numerical data and the methods used are approximate ones. Hence, the error in a computed result may be caused by the errors in the data or the errors in the method or both. Generally, the numbers are represented in **decimal** (base 10) form, while in computers the numbers are represented using the **binary** (base 2) and also the **hexadecimal** (base 16) forms. To perform a numerical calculation, approximate them first by a representation involving a finite number of **significant digits**. If the numbers to be represented are very large or very small, then they are written in **floating point notation**. The Institute of Electrical and Electronics Engineers (IEEE) has published a standard for binary floating point arithmetic. This standard, known as the IEEE Standard 754, has been widely adopted. The standard specifies formats for **single precision** and **double precision** numbers. The simplest way of reducing the number of significant digits in the representation of a number is simply to ignore the unwanted digits known as chopping. All these topics are discussed in the following section.

### Significant Figures

In approximate representation of numbers, the number is represented with a finite number of digits. All the digits in the usual decimal representation may not be significant while considering the accuracy of the number. Consider the following numbers:

$$1514,\ 15.14,\ 1.324,\ 1524$$

Each of them has four significant digits and all the digits in them are significant. Now consider the following numbers,

$$0.00215,\ 0.0215,\ 0.000215,\ 0.0000125$$

The leading zeroes after the decimal point in each of the above numbers are not significant. Each number has only three significant digits, even though they have different number of digits after the decimal point.

### Floating Point Computation

Every real number is usually represented by a finite or infinite sequence of decimal digits. This is called decimal system representation. For example, we can represent $\frac{1}{4}$ as 0.25, but $\frac{1}{3}$ as 0.333... Thus $\frac{1}{4}$ is represented by two significant digits only, while $\frac{1}{3}$ is represented by an infinite number of digits. Most computers have two

NOTES

forms of storing numbers for performing computations. They are fixed-point and floating point. In a fixed-point system, all numbers are given with a fixed number of decimal places. For example, 35.123, 0.014, 2.001. However, fixed-point representation is not of practical importance in scientific computation, since it cannot deal with very large or very small numbers.

In a floating-point representation, a number is represented with a finite number of significant digits having a floating decimal point. We can express the floating decimal number as follows:

$$623.8 \text{ as } 0.6238 \times 10^3,\ 0.0001714 \text{ as } 0.1714 \times 10^{-3}$$

A very large number can also be representated with floating-point representation, keeping the first few significant digits such as $0.14263218 \times 10^{39}$. Similarly, a very small number can be written with only the significant digits, leaving the leading zeros such as $0.32192516 \times 10^{-19}$.

In the decimal system, very large and very small numbers are expressed in scientific notation as follows: $4 \cdot 69 \times 10^{23}$ and $1 \cdot 601 \times 10^{-19}$. Binary numbers can also be expressed by the floating point representation. The floating point representation of a number consists of two parts: the first part represents a signed, fixed point number called the *mantissa* (*m*); the second part designates the position of the decimal (or binary) point and is called the *exponent* (*e*). The fixed point mantissa may be a fraction or an integer. The number of bits required to express the exponent and mantissa is determined by the accuracy desired from the computing system as well as its capability to handle such numbers. For example, the decimal number + 6132.789 is represented in floating point as follows:

$$\underbrace{\overset{\text{sign}}{0} \quad 0 \cdot 6132789}_{\text{mantissa}} \quad \underbrace{\overset{\text{sign}}{0} \quad 04}_{\text{exponent}}$$

The mantissa has a 0 in the leftmost position to denote a plus. Here, the mantissa is considered to be a fixed point fraction. This representation is equivalent to the number expressed as a fraction 10 times by an exponent, that is $0.6132789 \times 10^{+04}$. Because of this analogy, the mantissa is sometimes called the *fraction part*.

Consider, for example, a computer that assumes integer representation for the mantissa and radix 8 for the numbers. The octal number $+ 36.754 = 36754 \times 8^{-3}$ in its floating point representation will look like this:

$$\underbrace{\overset{\text{sign}}{0} \quad 36754}_{\text{mantissa}} \quad \underbrace{\overset{\text{sign}}{1} \quad 03}_{\text{exponent}}$$

When this number is represented in a register in its binary-coded form, the actual value of the register becomes 0 011 110 111 101 100 and 1 000 011.

NOTES

Most computers and all electronic calculators have a built-in capacity to perform floating-point arithmetic operations.

**Example 1:** Determine the number of bits required to represent in floating point notation the exponent for decimal numbers in the range of $10^{\pm 86}$.

**Solution:** Let $n$ be the required number of bits to represent the number $10^{\pm 86}$.

$$2^n = 10^{86}$$

$$n \log 2 = 86$$

$$n = \frac{86}{\log 2} = \frac{86}{0.3010} = 285.7$$

Therefore, $10^{\pm 86} = 2^{\pm 285.7}$

The exponent ±285 can be represented by a 10-bit binary word. It has a range of exponents (+511 to –512).

### Errors in Numerical Solution

The errors in a numerical solution are basically of two types. They are *truncation error* and *computational error*. The error which is inherent in the numerical method employed for finding numerical solution is called the truncation error. The computational error arises while doing arithmetic computation due to representation of numbers with a finite number of decimal digits.

The truncation error arises due to the replacement of an infinite process such as summation or integration by a finite one. For example, in computation of a transcendental function we use Taylor series/Maclaurin series expansion but retain only a finite number of terms. Similarly, a definite integral is numerically evaluated using a finite sum with a few function values of the integral. Thus, we express the error in the solution obtained by numerical method.

Inherent errors are errors in the data which are obtained by physical measurement and are due to limitations of the measuring instrument. The analysis of errors in the computed result due to the inherent errors in data is similar to that of round-off errors.

### Generation and Propagation of Round-Off Error

During numerical computation on a computer, a round-off error is generated by taking an infinite decimal representation of a real, rational number such as 1/3, 4/7, etc., by a finite size decimal form. In each arithmetic operation with such approximate rounded-off numbers there arises a round-off error. Also round-off errors present in the data will propagate in the result. Consider two approximate floating point numbers rounded-off to four significant digits.

$$x = 0.2234 \times 10^3 \text{ and } y = 0.1112 \times 10^2$$

The sum $x + y = 0.23452 \times 10^3$ is rounded-off to $0.23456 \times 10^3$ with an absolute error, $2 \times 10^{-2}$. This is the new round-off error generated in the result. Besides this error, the result will have an error propagated from the round-off errors in $x$ and $y$.

NOTES

### Round-Off Errors in Arithmetic Operations

To get an insight into the propagation of round-off errors, let us consider them for the four basic operations of addition, subtraction, multiplication and division. Let $x_T$ and $y_T$ be two real numbers whose round-off errors in their approximate representations $x$ and $y$ are $\varepsilon_1$ and $\varepsilon_2$ respectively, so that

$$x_T = x + \varepsilon_1 \quad \text{and} \quad y_T = y + \varepsilon_2$$

Their addition gives, $(x_T + y_T) = (x + y) + \varepsilon_1 + \varepsilon_2$

Hence, the propagated round-off error is given by,

$$(x_T + y_T) - (x + y) = \varepsilon_1 + \varepsilon_2$$

Thus the propagated round-off error is the sum of two approximate numbers (having round-off errors) equal to the sum of the round-off errors in the individual numbers.

The multiplication of two approximate numbers has the propagated round-off error given by,

$$x_T \times y_T - xy = \varepsilon_1 y + \varepsilon_2 x + \varepsilon_1 \varepsilon_2$$

Since the product $\varepsilon_1\varepsilon_2$ is a small quantity of higher order, then $\varepsilon_1$ or $\varepsilon_2$ may take the propagated round-off error as $\varepsilon_1 x_1 + \varepsilon_2 y_1$ and the relative propagated error is given by,

$$\frac{\varepsilon_1 x + \varepsilon_2 y}{xy} = \frac{\varepsilon_1}{x} + \frac{\varepsilon_2}{y}$$

This is equal to the sum of the relative errors in the numbers $x$ and $y$.

Similarly, for division we get the relative propagated error as,

$$\frac{\frac{x_T}{y_T} - \frac{x}{y}}{\frac{x}{y}} = \frac{\varepsilon_1}{x} - \frac{\varepsilon_2}{y}$$

Thus, the relative error in division is equal to the difference of the relative errors in the numbers.

### Errors in Evaluation of Functions

The propagated error in the evaluation of a function $f(x)$ of a single variable $x$ having a round-off error $\varepsilon$ is given by,

$$f(x + \varepsilon) - f(x) \approx \varepsilon \quad f'(x)$$

NOTES

In the evaluation of a function of several variables $x_1, x_2, \ldots, x_n$, the propagated round-off error is given by $\sum_{i=1}^{n} \varepsilon_1 \frac{\partial f}{\partial x_i}$, where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are the round-off errors in $x_1, x_2, \ldots, x_n$, respectively.

**Significance Errors**

During arithmetic computations of approximate numbers having fixed precision, there may be loss of significant digits in some cases. The error due to loss of significant digits is termed as significance error. Significance error is more serious than round-off errors, since it affects the accuracy of the result.

There are two situations when loss of significant digits occur. These are,

(*i*) Subtraction of two nearly equal numbers.

(*ii*) Division by a very small divisor compared to the dividend.

For example, consider the subtraction of the nearly equal numbers $x = 0.12454657$ and $y = 0.12452413$, each having eight significant digits. The result $x - y = 0.22440000 \times 10^{-4}$, is correct to four significant figures only. This result when used in further computations leads to serious error in the result.

Consider the problem of computing the roots of the quadratic equation,

$$ax^2 + bx + c = 0$$

The roots of this equation are,

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

If $b^2 >> 4ac$, then the evaluation of $-b + \sqrt{b^2 - 4ac}$ leads to subtraction of nearly equal numbers. One can avoid this by rewriting the expression,

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

It can be written as,

$$\frac{\left(-b + \sqrt{b^2 - 4ac}\right)\left(-b - \sqrt{b^2 - 4ac}\right)}{2a \times \left(b - \sqrt{b^2 - 4ac}\right)} = \frac{-2c}{b + \sqrt{b^2 - 4ac}}$$

Let the quadratic equation be,

$$x^2 + 100.0001x + 0.01 = 0$$

Using the first formula, we get the smaller root $= 0.10050000 \times 10^{-3}$, whereas exact root is $0.10000\ 0000 \times 10^{-3}$. But using the last expression we get the smaller root as $0.10000000 \times 10^{-3}$ which does not has the effect of significance error.

Consider an example where loss of significant digits occur due to division by a small number.

Computation of $f(x) = \frac{1 - \cos x}{x^2}$, for small values of $x$ would have loss of significant digits.

The Table 7.1 shows the computed values of $f(x)$ upto six decimal places along with the correct values and error.

***Table 7.1** Computed Value of f(x) upto Six Decimal Places*

| $x$ | Computed $f(x)$ | Correct $f(x)$ | Error |
|---|---|---|---|
| 0.1 | 0.499584 | 0.499583 | – 0.000001 |
| 0.01 | 0.50008 | 0.499996 | – 0.000012 |
| 0.001 | 0.506639 | 0.500000 | – 0.006639 |
| 0.0001 | 0.500000 | 0.745058 | 0.245058 |

Table 7.1 shows that the error in the computed value becomes more serious for smaller value of $x$. It may be noted that the correct values of $f(x)$ can be computed by avoiding the divisions by small number by rewriting $f(x)$ as given below.

$$f(x) = \frac{1 - \cos x}{x^2} \times \frac{1 + \cos x}{1 + \cos x}$$

i.e.,
$$f(x) = \frac{\sin^2 x}{x^2 (1 + \cos x)}$$

### Characteristics of Numerical Computation

A numerical solution can never be exact but attempts are made to know the accuracy of the approximate solution. Thus one attempts to get an approximate solution which differs from the exact solution by less than a specified tolerance limit.

Some numerical methods find the solution by a direct method but many others are of repetitive nature. The first step in the solution procedure is to take an approximate solution. Then the numerical method is applied repeatedly to get better results till the solution is obtained up to a desired accuracy. This process is known as iteration.

To get a numerical solution on a computer, one has to write an algorithm. An algorithm is a sequence of unambiguous steps used to solve a given problem. In the design of such computer programs one considers the input data required to implement the numerical method and writes the computer program in a suitable programming language. The output of the program should give the solution with the desired accuracy.

It may be noted that the iterative method gives rise to a sequence of results. The convergence of this sequence to get the output upto a desired accuracy is dependent on the initial data. Hence, one has to suitably choose the input data. Thus, if for some input data the sequence is not convergent for certain pre-assigned

NOTES

number of iterations then the input data is changed. It is for this reason that one has to limit the number of iterations to be employed while designing the computer program.

While computing a solution with the help of an algorithm, one has to check the correctness of the solution obtained. To do so, one has to have some test data whose solution is known.

**Example 2:** The numbers 28.483 and 27.984 are both approximate and are correct up to the last digits shown. Compute their difference. Indicate how many significant digits are present in the result and comment.

**Solution:** We have 28.483 – 27.984 = 00.499. The result has only three significant digits. This is due to the loss of significant digits during subtraction of nearly equal numbers.

**Example 3:** Round the number $x = 2.2554$ to three significant figures. Find the absolute error and the relative error.

**Solution:** The rounded-off number is 2.25.

The absolute error is 0.0054.

The relative error is $\simeq \frac{0.0054}{2.25} = 0.0024$

The percentage error is 0.24 per cent.

**Example 4:** If $\pi = 3.14$ instead of $\frac{22}{7}$, find the relative error.

**Solution:** Relative error $= \left(\frac{22}{7} - 3.14\right) \Big/ \frac{22}{7} = 0.00090.$

**Example 5:** Determine the number of correct digits in $x = 0.2217$, if it has a relative error $\varepsilon_r = 0.2 \times 10^{-1}$.

**Solution:** Absolute error = $0.2 \times 10^{-1} \times 0.2217 = 0.004434$

Hence, $x$ has only one correct digit $x \simeq 0.2$.

**Example 6:** Round-off the number 4.5126 to four significant figures and find the relative percentage error.

**Solution:** The number 4.5126 rounded-off to four significant figures is 4.153.

Relative error $= \frac{-\ 0.0004}{4.5126} \times 100 = -\ 0.0088$ per cent

**Example 7:** Given $f(x, y, z) = \frac{5xy^2}{z^2}$, find the relative maximum error in the evaluation of $f(x, y, z)$ at $x = y = z = 1$, if $x, y, z$ have absolute errors $\Delta x = \Delta y = \Delta z = 0.1$

**Solution:** The value of $f(x, y, z)$ at $x = y = z = 1$ is 5. The maximum absolute error in the evaluation of $f(x, y, z)$ is,

$$|(\Delta f)_{\max}| = \left|\frac{\partial f}{\partial x}\Delta x\right| + \left|\frac{\partial f}{\partial y}\Delta y\right| + \left|\frac{\partial f}{\partial z}\Delta z\right|$$

$$= \left|\frac{5y^2}{z^2}\Delta x\right| + \left|\frac{10xy}{z^2}\Delta y\right| + \left|\frac{-10xy^2}{z^3}\Delta z\right|$$

At, $x = y = z = 1$, the maximum relative error is,

$$(E_R)_{\max} = \frac{25 \times 0.1}{5} = 0.5$$

**Example 8:** Find the relative propagated error in the evaluation of $x + y$ where $x = 13.24$ and $y = 14.32$ have round-off errors $\varepsilon_1 = 0.004$ and $\varepsilon_2 = 0.002$ respectively.

**Solution:** Here, $x + y = 27.56$ and $\varepsilon_1 + \varepsilon_2 = 0.006$.

Thus, the required relative error $= \frac{0.006}{27.56} = 0.0002177$.

**Example 9:** Find the relative percentage error in the evaluation of $u = xy$ with $x = 5.43$, $y = 3.82$ having round-off errors 0.01 in both $x$ and $y$.

**Solution:** Now, $xy = 5.43 \times 3.82 \simeq 20.74$

The relative error in $x$ is $\frac{0.01}{5.43} \simeq 0.0018.$

The relative error in $y$ is $\frac{0.01}{3.82} \simeq 0.0026.$

Thus, the relative propagated error in $x$ and $y = 0.0044$.

The percentage relative error = 0.44 per cent.

**Example 10:** Given $u = xy + yz + zx$, find the estimate of relative percentage error in the evaluation of $u$ for $x = 2.104$, $y = 1.935$ and $z = 0.845$. What are the approximate values correct to the last digit?

**Solution:** Here, $u = x(y + z) + yz = 2.104\,(1.935 + 0.845) + 1.935 \times 0.845$

$$= 5.849 + 1.635 = 7.484$$

$$\text{Error, } \Delta u = (y+z)\Delta x + (z+x)\Delta y + (x+y)\Delta z$$

$$= 0.0005 \times 2(x+y+z) \qquad (\because\ \Delta x = \Delta y = \Delta z = 0.0005)$$

$$\simeq 2 \times 4.884 \times 0.0005 \simeq 0.0049$$

Hence, the relative percentage error $= \frac{0.0049}{7.884} \times 100 = 0.062$ per cent.

**Example 11:** The diameter of a circle measured to within 1 mm is $d = 0.842$ m. Compute the area of the circle and give the estimated relative error in the computed result.

NOTES

**Solution:** The area of the circle $A$ is given by the formula, $A = \frac{\pi d^2}{4}$.

NOTES

Thus, $A = \frac{3.1416}{4} \times (0.842)^2$ m$^2$ = 0.5568 m$^2$.

Here the value of $\pi$ is taken upto 4th decimal palce since the data of $d$ has accuracy upto the 3rd decimal place. Now the relative percentage error in the above computation is,

$$E_p = \frac{2\pi d}{4} \times \frac{4\Delta d}{\pi d^2} \times 100 = \frac{2\Delta d}{d} \times 100 = \frac{2 \times 0.01}{0.842} = 0.24 \text{ per cent}$$

**Example 12:** The length $a$ and the width $b$ of a plate is measured accurate up to 1cm as $a = 5.43$ m and $b = 3.82$ m. Compute the area of the plate and indicate its error.

**Solution:** The area of the plate is given by,

$A = ab = 3.82 \times 5.43$ sq. m. = 20.74 m$^2$.

The estimate of error in the computed value of $A$ is given by,

$$\begin{aligned} \Delta A &= \Delta a . b + \Delta b . a \\ &= 0.01 \times 3.82 + 0.01 \times 5.43, \quad \text{since} \quad \Delta a = \Delta b = 0.01 \\ &= 0.0925 \simeq 10 \text{ m}^2 \end{aligned}$$

**Computational Algorithms**

For solving problems with the help of a computer, one should first analyse the mathematical formulation of the problem and consider a suitable numerical method for solving it. The next step is to write an algorithm for implementing the method. An algorithm is defined as a finite sequence of unambiguous steps to be followed for solving a given problem. Finally, one has to write a computer program in a suitable programming language. A computer program is a sequence of computer instructions for solving a problem.

It is possible to write more than one algorithm to solve a specific problem. But one should analyse them before writing a computer program. The analysis involves checking their correctness, robustness, efficiency and other characteristics. The analysis is helpful for solving the problem on a computer. The analysis of correctness of an algorithm ensures that the algorithm gives a correct solution of the problem. The analysis of robustness is required to ascertain if the algorithm is capable of tackling the problem for possible cases or for all possible variations of the parameters of the problem. The efficiency is concerned with the computational complexities and the total time required to solve the problem.

Computer oriented numerical methods must deal with algorithms for implementation of numerical methods on a computer. The following algorithms of some simple problems will make the concept clear.

NOTES

Consider the problem of solving a pair of linear equations in two unknowns given by,

$$a_1x + b_1y = c_1$$
$$a_2x + b_2y = c_2$$

where $a_1, b_1, c_1, a_2, b_2, c_2$ are real constants. The solution of the equations are given by cross multiplication as,

$$x = \frac{b_2c_1 - b_1c_2}{a_1b_2 - a_2b_1}, \qquad y = \frac{c_2a_1 - c_1a_2}{a_1b_2 - a_2b_1}$$

It may be noted that if $a_1 b_2 - a_2 b_1 = 0$, then the solution does not exist. This aspect has to be kept in mind while writing the algorithm as given below.

**Algorithm:** Solution of a pair of equations $a_1x + b_1y = c_1$, $a_2 \ x + b_2 \ y = c_2$

**Step 1:** Read $a_1, b_1, c_1, a_2, b_2, c_2$

**Step 2:** Compute $d = a_1 b_2 - a_2 b_1$

**Step 3:** Check if $d = 0$, then go to Step 8 else go to next step

**Step 4:** Compute $x$

**Step 5:** Compute $y$

**Step 6:** Write '$x$ =', $x$, '$y$ =', $y$

**Step 7:** Go to Step 9

**Step 8:** Write 'no solution'

**Step 9:** Stop

**Example 13:** Write an algorithm to compute the roots of a quadratic equation, $ax^2 + bx + c = 0$.

**Solution:** We know that the roots of the quadratic equation are given by,

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Further, if $b^2 \geq 4ac$, the roots are real, otherwise they are complex conjugates. This aspect is to be considered while writing an algorithm.

**Algorithm:** Computation of roots of a quadratic equation.

**Step 1:** Read $a, b, c$

**Step 2:** Compute $d = b^2 - 4ac$

**Step 3:** Check if $d \geq 0$, go to Step 4 else go to Step 8

**Step 4:** Compute $x_1 = (-b + \sqrt{d})/(2a)$

**Step 5:** Compute $x_2 = (-b - \sqrt{d})/(2a)$

**Step 6:** Write 'Roots are real', $x_1, x_2$

**Step 7:** Go to Step 11

**Step 8:** Compute $x_i = \sqrt{-d}/(2a)$

NOTES

**Step 9:** Compute $x_r = -b/(2a)$

**Step 10:** Write 'Roots are complex', 'Real part =' $x_r$ 'Imaginary part =', $x_i$

**Step 11:** Stop

**Check Your Progress**

1. What are the two parts of floating point representation?
2. Define truncation and computational errors.
3. Define inherent errors.
4. What is propagated round-off error?
5. What are significance errors?
6. Write the situations when loss of significant digits occur.
7. Why we write an algorithm?
8. Define features and purpose of computational algorithms.

## 7.3 LEAST SQUARE APPROXIMATION

In this section, we consider the problem of approximating an unknown function whose values, at a set of points, are generally known only empirically and are, thus subject to inherent errors, which may sometimes be appreciably larger in many engineering and scientific problems. In these cases, it is required to derive a functional relationship using certain experimentally observed data. Here the observed data may have inherent or round-off errors, which are serious, making polynomial interpolation for approximating the function inappropriate. In polynomial interpolation the truncation error in the approximation is considered to be important. But when the data contains round-off errors or inherent errors, interpolation is not appropriate.

The subject of this section is curve fitting by least square approximation. Here we consider a technique by which noisy function values are used to generate a smooth approximation. This smooth approximation can then be used to approximate the derivative more accurately than with exact polynomial interpolation.

There are situations where interpolation for approximating function may not be efficacious procedure. Errors will arise when the function values $f(x_i)$, $i = 1, 2,$

NOTES

…, $n$ are observed data and not exact. In this case, if we use the polynomial interpolation, then it would reproduce all the errors of observation. In such situations one may take a large number of observed data, so that statistical laws in effect cancel the errors introduced by inaccuracies in the measuring equipment. The approximating function is then derived, such that the sum of the squared deviation between the observed values and the estimated values are made as small as possible.

Mathematically, the problem of curve fitting or function approximation may be stated as follows:

To find a functional relationship $y = g(x)$, that relates the set of observed data values $P_i(x_i, y_i)$, $i = 1, 2, ..., n$ as closely as possible, so that the graph of $y = g(x)$ goes near the data points $P_i$'s though not necessarily through all of them.

The first task in curve fitting is to select a proper form of an approximating function $g(x)$, containing some parameters, which are then determined by minimizing the total squared deviation.

For example, $g(x)$ may be a polynomial of some degree or an exponential or logarithmic function. Thus $g(x)$ may be any of the following:

(*i*) $g(x) = \alpha + \beta x$ (*ii*) $g(x) = \alpha + \beta x + \gamma x^2$

(*iii*) $g(x) = \alpha e^{\beta x}$ (*iv*) $g(x) = \alpha e^{-\beta x}$

(*v*) $g(x) = \alpha \log(\beta x)$

Here $\alpha, \beta, \gamma$ are parameters which are to be evaluated so that the curve $y = g(x)$, fits the data well. A measure of how well the curve fits is called the goodness of fit.

In the case of least square fit, the parameters are evaluated by solving a system of normal equations, derived from the conditions to be satisfied so that the sum of the squared deviations of the estimated values from the observed values, is minimum.

### Method of Least Squares

Let $(x_1, f_1), (x_2, f_2), ..., (x_n, f_n)$ be a set of observed values and $g(x)$ be the approximating function. We form the sums of the squares of the deviations of the observed values $f_i$ from the estimated values $g(x_i)$,

i.e., $$S = \sum_{i=1}^{n} \{f_i - g(x_i)\}^2 \quad (7.1)$$

The function $g(x)$ may have some parameters, $\alpha, \beta, \gamma$. In order to determine these parameters we have to form the necessary conditions for $S$ to be minimum, which are:

$$\frac{\partial S}{\partial \alpha} = 0, \frac{\partial S}{\partial \beta} = 0, \frac{\partial S}{\partial \gamma} = 0 \qquad (7.2)$$

NOTES

These equations are called normal equations, solving which we get the parameters for the best approximate function $g(x)$.

**Curve Fitting by a Straight Line:** Let $g(x) = \alpha + \beta\, x$, be the straight line which fits a set of observed data points $(x_i, y_i)$, $i = 1, 2, ..., n$.

Let $S$ be the sum of the squares of the deviations $g(x_i) - y_i$, $i = 1, 2,...,n$, given by,

$$S = \sum_{i=1}^{n} (\alpha + \beta\, x_i - y_i)^2 \qquad (7.3)$$

We now employ the method of least squares to determine $\alpha$ and $\beta$ so that $S$ will be minimum. The normal equations are,

$$\frac{\partial S}{\partial \alpha} = 0, \text{ i.e., } \sum_{i=1}^{n} (\alpha + \beta\, x_i - y_i) = 0 \qquad (7.4)$$

And,
$$\frac{\partial S}{\partial \beta} = 0, \text{ i.e., } \sum_{i=1}^{n} x_i (\alpha + \beta\, x_i - y_i) = 0 \qquad (7.5)$$

These conditions give,

$$n\alpha + S_1\beta - S_{01} = 0$$
$$S_1\alpha + S_2\beta - S_{11} = 0$$

Where,
$$S_1 = \sum_{i=1}^{n} x_i, \quad S_{01} = \sum_{i=1}^{n} y_i, \quad S_2 = \sum_{i=1}^{n} x_i^2, \quad S_{11} = \sum_{i=1}^{n} x_i y_i$$

Solving,

$$\frac{\alpha}{-S_1 S_{11} + S_1 S_2} = \frac{\beta}{nS_{11} - S_1 S_{01}} = \frac{1}{nS_2 - S_1^2}. \quad \text{Also} \quad \alpha = \frac{S_{01}}{n} - \beta \frac{S_1}{n}.$$

**Algorithm:** Fitting a straight line $y = a + bx$.

**Step 1:** Read $n$ [$n$ being the number of data points]

**Step 2:** Initialize : sum $x = 0$, sum $x^2 = 0$, sum $y = 0$, sum $xy = 0$

**Step 3:** For $j = 1$ to $n$ compute

Begin

Read data $x_j$, $y_j$

Compute sum $x$ = sum $x + x_j$

Compute sum $x^2 + x_j \times x_j$

Compute sum $y$ = sum $y + y_i \times y_j$

Compute sum $xy$ = sum $xy + x_j \times y_j$

End

**Step 4:** Compute $b = (n \times \text{sum } xy - \text{sum } x \times \text{sum } y)/ (n \times \text{sum } x^2 - (\text{sum } x)^2)$

**Step 5:** Compute $x$ bar = sum $x$ / $n$

**Step 6:** Compute $y$ bar = sum $y$ / $n$

**Step 8:** Compute $a = y$ bar – $b \times x$ bar

**Step 9:** Write $a$, $b$

**Step 10:** For $j = 1$ to $n$

Begin

Compute $y$ estimate = $a + b \times x$

write $x_j, y_j, y$ estimate

End

**Step 11:** Stop

NOTES

**Curve Fitting by a Quadratic (A Parabola):** Let $g(x) = a + bx + cx^2$, be the approximating quadratic to fit a set of data $(x_i, y_i)$, $i = 1, 2, ..., n$. Here the parameters are to be determined by the method of least squares, i.e., by minimizing the sum of the squares of the deviations given by,

$$S = \sum_{i=1}^{n} (a + bx_i + cx_i^2 - y_i)^2 \tag{7.6}$$

Thus the normal equations, $\frac{\partial S}{\partial a} = 0$, $\frac{\partial S}{\partial b} = 0$, $\frac{\partial S}{\partial c} = 0$, are as follows: (7.7)

$$\sum_{i=1}^{n} (a + bx_i + cx_i^2 - y_i) = 0$$
$$\sum_{i=1}^{n} x_i (a + bx_i + cx_i^2 - y_i) = 0$$
$$\sum_{i=1}^{n} x_i^2 (a + bx_i + cx_i^2 - y_i) = 0. \tag{7.8}$$

These equations can be rewritten as,

$$na + s_1 b + s_2 c - s_{01} = 0$$
$$s_1 a + s_2 b + s_3 c - s_{11} = 0$$
$$s_2 a + s_3 b + s_4 c - s_{21} = 0 \tag{7.9}$$

where $s_1 = \sum_{i=1}^{n} x_i$, $s_2 = \sum_{i=1}^{n} x_i^2$, $s_3 = \sum_{i=1}^{n} x_i^3$, $s_4 = \sum_{i=1}^{n} x_i^4$

$$s_{01} = \sum_{i=1}^{n} y_i, \quad s_{11} = \sum_{i=1}^{n} x_i \ y_i, \quad s_{21} = \sum_{i=1}^{n} x_i^2 y_i \tag{7.10}$$

NOTES

It is clear that the normal equations form a system of linear equations in the unknown parameters $a$, $b$, $c$. The computation of the coefficients of the normal equations can be made in a tabular form for desk computations as shown below.

| | $x$ | $x_i$ | $y_i$ | $x_i^2$ | $x_i^3$ | $x_i^4$ | $x_i y_i$ | $x_i^2 y_i$ |
|---|---|---|---|---|---|---|---|---|
| | 1 | $x_1$ | $y_1$ | $x_1^2$ | $x_1^3$ | $x_1^4$ | $x_1 y_1$ | $x_1^2 y_1$ |
| | 2 | $x_2$ | $y_2$ | $x_2^2$ | $x_2^3$ | $x_2^4$ | $x_2 y_2$ | $x_2^2 y_2$ |
| | ... | ... | ... | ... | ... | ... | ... | ... |
| | $n$ | $x_n$ | $y_n$ | $x_n^2$ | $x_n^3$ | $x_n^4$ | $x_n y_n$ | $x_n^2 y_n$ |
| Sum | | $s_1$ | $s_{01}$ | $s_2$ | $s_3$ | $s_4$ | $s_{11}$ | $s_{21}$ |

The system of linear equations can be solved by Gaussian elimination method. It may be noted that number of normal equations is equal to the number of unknown parameters.

**Example 14:** Find the straight line fitting the following data:

| $x_i$ | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|
| $y_1$ | 13.72 | 12.90 | 12.01 | 11.14 | 10.31 |

**Solution:** Let $y = a + bx$, be the straight line which fits the data. We have the normal equations $\frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b} = 0$ for determining $a$ and $b$, where $S = \sum_{i=1}^{5} (y_i - a - bx_i)^2$ .

Thus, $$\sum_{i=1}^{5} y_i - na - b\sum_{i=1}^{5} x_i = 0$$

and , $$\sum_{i=1}^{5} x_i y_i - a\sum_{I=1}^{5} x_i - b\sum_{i=1}^{5} x_i^2 = 0$$

The coefficients are computed in the table below.

| | $x_i$ | $y_i$ | $x_i^2$ | $x_i y_i$ |
|---|---|---|---|---|
| | 4 | 13.72 | 16 | 54.88 |
| | 6 | 12.90 | 36 | 77.40 |
| | 8 | 12.01 | 64 | 96.08 |
| | 10 | 11.14 | 100 | 111.40 |
| | 12 | 10.31 | 144 | 123.72 |
| Sum | 40 | 60.08 | 360 | 463.48 |

Thus the normal equations are,

$$5a + 40b - 60.08 = 0$$
$$40a + 360b - 463.48 = 0$$

Solving these two equations we obtain,

$$a = 15.448,\ b = 0.429$$

Thus, $y = g(x) = 15.448 - 0.429\,x$, is the straight line fitting the data.

NOTES

**Example 15:** Use the method of least square approximation to fit a straight line to the following observed data:

| | | | | | |
|---|---|---|---|---|---|
| $x_i$ | 60 | 61 | 62 | 63 | 64 |
| $y_i$ | 40 | 40 | 48 | 52 | 55 |

**Solution:** Let the straight line fitting the data be $y = a + bx$. The data values being large, we can use a change in variable by substituting $u = x - 62$ and $v = y - 48$.

Let $v = A + B\,u$, be a straight line fitting the transformed data, where the normal equations for $A$ and $B$ are,

$$\sum_{i=1}^{5} v_i = 5A + B\sum_{i=1}^{5} u_i$$
$$\sum_{i=1}^{5} u_i v_i = A\sum_{i=1}^{5} u_i + B\sum_{i=1}^{5} u_i^2$$

The computation of the various sums are given in the table below,

| | $x_i$ | $y_i$ | $u_i$ | $v_i$ | $u_i v_i$ | $u_i^2$ |
|---|---|---|---|---|---|---|
| | 60 | 40 | –2 | –8 | 16 | 4 |
| | 61 | 42 | –1 | –6 | 6 | 1 |
| | 62 | 48 | 0 | 0 | 0 | 0 |
| | 63 | 52 | 1 | 4 | 4 | 1 |
| | 64 | 55 | 2 | 7 | 14 | 4 |
| Sum | | | 0 | –3 | 40 | 10 |

Thus the normal equations are,

$$-3 = 5A \quad \text{and} \quad 40 = 10B$$
$$\therefore A = -\frac{3}{5} \quad \text{and} \quad B = 4$$

This gives the line, $v = -3/5 + 4u$

Or, $$20u - 5v - 3 = 0.$$

Transforming we get the line,

$$20\,(x - 62) - 5\,(y - 48) - 3 = 0$$

Or, $$20\,x - 5y - 1003 = 0$$

**Curve Fitting with an Exponential Curve:** We consider a two parameter exponential curve as,

$$y = ae^{-bx} \qquad (7.11)$$

NOTES

For determining the parameters, we can apply the principle of least squares by first using a transformation,

$z = \log y$, so that Equation (7.11) is rewritten as, (7.12)

$$z = \log a - bx \qquad (7.13)$$

Thus, we have to fit a linear curve of the form $z = \alpha + \beta \quad x$, with $z-x$ variables and then get the parameters $a$ and $b$ as,

$$a = e^{\alpha}, \quad b = -\beta \qquad (7.14)$$

Thus proceeding as in linear curve fitting,

$$\beta = \frac{n\sum_{i=1}^{n} x_i \log y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \log y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (7.15)$$

And, $\alpha = \bar{z} - p\bar{x}$, where $\bar{x} = \left(\sum_{i=1}^{n} x_i\right) \Big/ n,\ \bar{z} = \left(\sum_{i=1}^{n} \log y_i\right) \Big/ n$ (7.16)

After computing $\alpha$ and $\beta$, we can determine $a$ and $b$ given by Equation (7.13). Finally, the exponential curve fitting data set is given by Equation (7.11).

**Algorithm:** To fit a straight line for a given set of data points by least square error method.

**Step 1:** Read the number of data points, i.e., $n$

**Step 2:** Read values of data-points, i.e., Read $(x_i, y_i)$ for $i = 1, 2,..., n$

**Step 3:** Initialize the sums to be computed for the normal equations,

i.e., $sx = 0$, $sx^2 = 0$, $sy = 0$, $syx = 0$

**Step 4:** Compute the sums, i.e., For $i = 1$ to $n$ do

Begin

$$sx = sx + x_i$$
$$sx^2 = sx^2 + x_i^2$$
$$sy = sy + y_i$$
$$syx = syx + x_i y_i$$

End

**Step 5:** Solve the normal equations, i.e., solve for $a$ and $b$ of the line $y = a + bx$

NOTES

Compute $d = n * sx^2 - sx * sx$

$b = (n * sxy - sy * sx) / d$

$x\text{bar} = sx / n$

$y\text{bar} = sy / n$

$a = y\text{bar} - b * x\,\text{bar}$

**Step 6:** Print values of $a$ and $b$

**Step 7:** Print a table of values of

$x_i, \quad y_i, \quad y_{pi} = a + bx_i \quad \text{for} \quad i = 1, \ 2, \ \ldots, \ n$

**Step 8:** Stop

**Algorithm:** To fit a parabola $y = a + bx + cx^2$, for a given set of data points by least square error method.

**Step 1:** Read $n$, the number of data points

**Step 2:** Read $(x_i, y_i)$ for $i = 1, 2, ..., n,$ the values of data points

**Step 3:** Initialize the sum to be computed for the normal equations,

i.e., $sx = 0$, $sx^2 = 0$, $sx^3 = 0$, $sx^4 = 0$, $sy = 0$, $sxy = 0$.

**Step 4:** Compute the sums, i.e., For $i = 1$ to $n$ do

Begin

$sx = sx + x_i$

$x^2 = x_i * x_i$

$sx^2 = sx^2 + x^2$

$sx^3 = sx^3 + x_i * x^2$

$sx^4 = sx^4 + x^2 * x^2$

$sy = sy + y_i$

$sxy = sxy + x_i * y_i$

$sx^2 y = sx^2 y + x^2 * y_i$

End

**Step 5:** Form the coefficients $\{a_{ij}\}$ matrix of the normal equations, i.e.,

$a_{11} = n, \quad a_{21} = sx, \quad a_{31} = sx^2$

$a_{12} = sx, \quad a_{22} = sx^2, \quad a_{32} = sx^3$

$a_{13} = sx^2, \quad a_{23} = sx^3, \quad a_{33} = sx^4$

**Step 6:** Form the constant vector of the normal equations.

$b_1 = sy$, $b_2 = sxy$, $b_3 = sx^2y$

**Step 7:** Solve the normal equation by Gauss-Jordan method

$$a_{12} = a_{12} / a_{11},\ a_{13} = a_{13} / a_{11},\ b_1 = b_1 / a_{11}$$
$$a_{22} = a_{22} - a_{21}a_{12},\ a_{23} = a_{23} - a_{21}a_{13}$$
$$b_2 = b_2 - b_1 a_{21}$$

NOTES

$$a_{32} = a_{32} - a_{31}a_{12}$$
$$a_{33} = a_{33} - a_{31}a_{13}$$
$$b_3 = b_3 - b_1 a_{31}$$
$$a_{23} = a_{23} / a_{22}$$
$$b_2 = b_2 / a_{22}$$
$$a_{33} = a_{33} - a_{23}\, a_{32}$$
$$b_3 = b_3 - a_{32}\, b_2$$
$$c = b_3 / a_{33}$$
$$b = b_2 - c\, a_{23}$$
$$a = b_1 - b\, a_{12} - c\, a_{13}$$

**Step 8:** Print values of $a, b, c$ (the coefficients of the parabola)

**Step 9:** Print the table of values of $x_k$, $y_k$ and $y_{pk}$ where $y_{pk} = a + bx_k + cx^2 k$,

i.e., print $x_k$, $y_k$, $y_{pk}$ for $k = 1, 2, ..., n$.

**Step 10:** Stop.

**Check Your Progress**

9. How is approximating function found in the method of least squares?
10. Explain the curve fitting by a straight line.

## 7.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The floating point representation of a number consists of mantissa and exponent.
2. The errors in a numerical solution are basically of two types. They are truncation error and computational error. The error which is inherent in the numerical method employed for finding numerical solution is called the truncation error. The computational error arises while doing arithmetic computation due to representation of numbers with a finite number of decimal digits.
3. Inherent errors are errors in the data which are obtained by physical measurement and are due to limitations of the measuring instrument. The analysis of errors in the computed result due to the inherent errors in data is similar to that of round-off errors.

4. Propagated round-off error is the sum of two approximate numbers (having round-off errors) equal to the sum of the round-off errors in the individual numbers.
5. During arithmetic computations of approximate numbers having fixed precision, there may be loss of significant digits in some cases. The error due to loss of significant digits is termed as significance error.
6. There are two situations when loss of significant digits occur. These are,
   (i) Subtraction of two nearly equal numbers
   (ii) Division by a very small divisor compared to the dividend
7. To get a numerical solution on a computer, one has to write an algorithm.
8. For solving problems with the help of a computer, one should first analyse the mathematical formulation of the problem and consider a suitable numerical method for solving it. The next step is to write an algorithm for implementing the method.
9. Let $(x_1,f_1)$, $(x_2,f_2)$,..., $(x_n,f_n)$ be a set of observed values and $g(x)$ be the approximating function. We form the sums of the squares of the deviations of the observed values $f_i$ from the estimated values $g(x_i)$,

   i.e., $S = \sum_{i=1}^{n} \{f_i - g(x_i)\}^2$

   The function $g(x)$ may have some parameters, $\alpha$, $\beta$, $\gamma$. In order to determine these parameters we have to form the necessary conditions for $S$ to be minimum, which are:

   $$\frac{\partial S}{\partial \alpha} = 0, \frac{\partial S}{\partial \beta} = 0, \frac{\partial S}{\partial \gamma} = 0$$

   These equations are called normal equations, solving which we get the parameters for the best approximate function $g(x)$.
10. Let $g(x) = \alpha + \beta\, x,$ be the straight line which fits a set of observed data points $(x_i, y_i)$, $i = 1, 2, ..., n$.
    Let $S$ be the sum of the squares of the deviations $g(x_i) - y_i$, $i = 1, 2,...,n$, given by,

    $$S = \sum_{i=1}^{n} (\alpha + \beta\, x_i - y_i)^2$$

## 7.5 SUMMARY

- Numerical methods are methods used for solving problems through numerical calculations providing a table of numbers and/or graphical representations or figures. Numerical methods emphasize that how the algorithms are implemented.

NOTES

- To perform a numerical calculation, approximate them first by a representation involving a finite number of significant digits. If the numbers to be represented are very large or very small, then they are written in floating point notation.
- The Institute of Electrical and Electronics Engineers (IEEE) has published a standard for binary floating point arithmetic.
- In approximate representation of numbers, the number is represented with a finite number of digits. All the digits in the usual decimal representation may not be significant while considering the accuracy of the number.
- In a floating representation, a number is represented with a finite number of significant digits having a floating decimal point.
- Floating point representation of a number consists of mantissa and exponent.
- The errors in a numerical solution are basically of two types termed as truncation error and computational error.
- The error which is inherent in the numerical method employed for finding numerical solution is called the truncation error.
- The truncation error arises due to the replacement of an infinite process such as summation or integration by a finite one.
- Inherent errors are errors in the data which are obtained by physical measurement and are due to limitations of the measuring instrument.
- The analysis of errors in the computed result due to the inherent errors in data is similar to that of round-off errors.
- Significance error is more serious than round-off errors.
- Iteration is the numerical method applied repeatedly to get better results till the solution is obtained up to a desired accuracy.
- An algorithm is a sequence of unambiguous steps used to solve a given problem.
- It is possible to write more than one algorithm to solve a specific problem.
- The algorithm analysis involves checking their correctness, robustness, efficiency and other characteristics.
- Single precision floating point format is a computer number format that occupies 4 bytes (32-bits) in computer memory and denotes wide range of values using a floating point.
- Double precision refers to a specific floating point number that has more precision, i.e., more digits to the right of the decimal point than a single precision number.

NOTES

## 7.6 KEY WORDS

- **Truncation error:** This error is inherent in the numerical method employed for finding numerical solution. It occurs due to the replacement of an infinite process such as summation or integration by a finite one.
- **Computational error:** This error occurs during arithmetic computation due to representation of numbers having a finite number of decimal digits.
- **Inherent error:** This error occurs in the data type which is obtained using physical measurement and also due to limitations of the measuring instruments.
- **Significance error:** This error occurs due to loss of significant digits.
- **Algorithm:** It is a sequence of finite steps used to solve a given problem.

## 7.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. What are floating point numbers?
2. Find the percentage error in approximating $\frac{5}{6}$ by $0.8333$ correct upto four significant figures.
3. Write the characteristics of numerical computation.
4. Find the relative error in the computation of $x - y$ for $x = 12.05$ and $y = 8.02$ having absolute errors $\Delta x = 0.005$ and $\Delta y = 0.001$.
5. Find the percentage error in computing $y = 3x^2 - 6x$ at $x = 1$, if the error in $x$ is 0.05.
6. Given $a = 1.135$ and $b = 1.075$ having absolute errors $\Delta a = 0.011$ and $\Delta b = 0.12$. Estimate the relative percentage error in the computation of $a - b$.
7. Find the percentage error in taking 1.33 as approximation for 4/3.
8. The length $a$ and breadth $b$ of a plate measured accurate to 1 cm as $a = 5.43$ m and $b = 3.82$ m. Estimate the area of the plate and estimate its absolute error.
9. How many significant digits are present in each of the following approximate numbers:

   10.54113, 5.4113, 0.054113, 0.00541

NOTES

## Long-Answer Questions

1. Round-off the following numbers to three decimal places:
   (*i*) 0.230582 (*ii*) 0.00221118 (*iii*) 2.3645 (*iv*) 1.3455
2. Round-off the following numbers to four significant figures:
   (*i*) 49.3628 (*ii*) 0.80022 (*iii*) 8.9325 (*iv*) 0.032588
   (*v*) 0.0029417 (*vi*) 0.00010211 (*vii*) 410.99
3. Round-off each of the following numbers to three significant figures and indicate the absolute error in each.
   (*i*) 49.3628 (*ii*) 0.9002 (*iii*) 8.325 (*iv*) 0.0039417
4. Find the sum of the following approximate numbers, correct to the last digits.
   0.348, 0.1834, 345.4, 235.2, 11.75, 0.0849, 0.0214, 0.0002435
5. Find the number of correct significant digits in the approximate number 11.2461. Given is its absolute error = $0.25 \times 10^{-2}$.
6. Given are the following approximate numbers with their relative errors. Determine the absolute errors.
   (*i*) $x_A = 12165, \quad \varepsilon_R = 0.1\%$ (*ii*) $x_A = 3.23, \quad \varepsilon_R = 0.6\%$
   (*iii*) $x_A = 0.798, \quad \varepsilon_R = 10\%$ (*iv*) $x_A = 67.84, \quad \varepsilon_R = 1\%$
7. Round-off the following numbers to four significant digits.
   (*i*) 450.92 (*ii*) 48.3668 (*iii*) 9.3265 (*iv*) 8.4155
   (*v*) 0.80012 (*vi*) 0.042514 (*vii*) 0.0049125
   (*viii*) 0.00020215
8. Write the following numbers in floating-point form rounded to four significant digits.
   (*i*) 100000 (*ii*) – 0.0022136 (*iii*) – 35.666
9. Determine the number of correct digits in the number *x* in each of the following (the relative errors are given).
   (*i*) $x = 0.2217, \quad \varepsilon_R = 0.2 \times 10^{-1}$ (*ii*) $x = 32.541, \quad \varepsilon_R = 0.1$
   (*iii*) $x = 0.12432, \quad \varepsilon_R = 10\%$ (*iv*) $x = 0.58632, \quad \varepsilon_R = 1\%$
10. Find the percentage error in computing $z = \sqrt{x}$ for $x = 4.44$, if $x$ is correct to its last digit only.
11. Let $u = 4x^6 + 3x - 9$. Find the relative percentage error in computing $u$ at $x = 1.1$, if the error in $x$ is 0.05.

NOTES

12. In the formula for computing $R$ for $R = \frac{r^2}{2h} + \frac{h}{2}$, find the absolute error in computing $R$ for $r = 48$ mm and $h = 56$ mm, due to errors of 1 mm in $r$ and 0.2 mm in $h$.
13. Find the smaller root of $0.001x^2 + 100.1x + 10000 = 0$, with the help of the usual formula and round-off to six significant digits. Compare with the correct answer $x = -100.0$.
14. Find the roots of the quadratic equation $x^2 - 100x - 0.1 = 0$, with the help of the usual formulae and show the significance error in the result.

## 7.8 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

NOTES

# UNIT 8 NUMERICAL INTEGRATION AND NUMERICAL DIFFERENTIATION

**Structure**



## 8.0 INTRODUCTION

Numerical integration methods can generally be described as combining evaluations of the integrand to get an approximation to the integral. The integrand is evaluated at a finite set of points called integration points and a weighted sum of these values is used to approximate the integral. The integration points and weights depend on the specific method used and the accuracy required from the approximation. Modern numerical integrations methods based on information theory have been developed to simulate information systems such as computer controlled systems, communication systems, and control systems.

In numerical analysis, numerical differentiation is the process of finding the numerical value of a derivative of a given function at a given point. It is the process of computing the derivatives of a function $f(x)$ when the function is not explicitly known, but the values of the function are known only at a given set of arguments $x = x_0, x_1, x_2, ..., x_n$. For finding the derivatives, a suitable interpolating polynomial is used and then its derivatives are used as the formulae for the derivatives of the function. Thus, for computing the derivatives at a point near the beginning of an equally spaced table, Newton's forward difference interpolation formula is used, whereas Newton's backward difference interpolation formula is used for computing the derivatives at a point near the end of the table.

In this unit, you will study about the numerical integration, numerical differentiation and extrapolation method.

## 8.1 OBJECTIVES

NOTES

After going through this unit, you will be able to:

- Understand the various numerical integrations
- Explain about the numerical differentiation
- Analyse the extrapolation method

## 8.2 NUMERICAL INTEGRATION

The evaluation of a definite integral cannot be carried out when the integrand $f(x)$ is not integrable, as well as when the function is not explicitly known but only the function values are known at a finite number of values of $x$. However, the value of the integral can be determined numerically by applying numerical methods. There are two types of numerical methods for evaluating a definite integral based on the following formula.

$$\int_a^b f(x)\, dx \qquad (8.1)$$

They are termed as Newton-Cotes quadrature and Gaussian quadrature. We first confine our attention to Newton-Cotes quadrature which is based on integrating polynomial interpolation formulae. This quadrature requires a table of values of the integrand at equally spaced values of the independent variable $x$.

### Newton-Cotes General Quadrature

We start with Newton's forward difference interpolation formula which uses a table of values of $f(x)$ at equally spaced points in the interval $[a, b]$. Let the interval $[a, b]$ be divided into $n$ equal sub-intervals such that,

$$a = x_0,\ x_i = x_o + ih, \text{ for } i = 1, 2, ..., n-1, \quad x_n = b \qquad (8.2)$$

so that, $nh = b–a$

Newton's forward difference interpolation formula is,

$$\varphi(s) = f_0 + s\ \Delta f_0 + \frac{s(s-1)}{2!}\Delta^2 f_0 + \frac{s(s-1)(s-2)}{3!}\Delta^3 f_0 + ... + \frac{s(s-1)(s-2)...(s-n+1)}{n!}\Delta^n f_0 \qquad (8.3)$$

Where $s = \dfrac{x - x_0}{h}$

Replacing $f(x)$ by $\phi(s)$ in Equation (8.1), we get

$$\int_{x_0}^{x_n} f(x)\ dx = h\int_0^n \left[ f_0 + s\Delta f_0 + \frac{s(s-1)}{2!}\Delta^2 f_0 + ... \right] ds$$

since when $x = x_0$, $s = 0$ and $x = x_n$, $s = n$ and $dx = h\ du$.

NOTES

Performing the integration on the RHS we have,

$$\int_{x_0}^{x_n} f(x)dx = h\left[nf_0 + \frac{n^2}{2}\Delta f_0 + \frac{1}{2}\left(\frac{n^3}{3} - \frac{n^2}{2}\right)\Delta^2 f_0 + \frac{1}{6}\left(\frac{n^4}{4} - 3\frac{n^3}{3} - 2\frac{n^2}{2}\right)\Delta^3 f_0 + \frac{1}{24}\left(\frac{n^5}{5} - \frac{3n^4}{2} + \frac{11n^3}{3} - 3n^2\right)\Delta^4 f_0 + \ldots\right] \quad (8.4)$$

We can derive different integration formulae by taking particular values of $n = 1, 2, 3, ....$ Again, on replacing the differences, the Newton-Cotes formula can be expressed in terms of the function values at $x_0, x_1,..., x_n$, as

$$\int_{x_0}^{x_n} f(x)dx = h\sum_{k=0}^{n} c_k f(x_k) \quad (8.5)$$

The error in the Newton-Cotes formula is given by,

$$E^n = \frac{h^{n+2}}{(n+1)!} f^{n+1} \quad (\xi) \cdot \int_0^n s(s-1)\cdots(s-n)ds \quad (8.6)$$

**Trapezoidal Formula of Numerical Integration**

Taking $n = 1$ in Equation (8.4), we get the trapezoidal formula given by,

$$\int_{x_0}^{x_1} f(x)dx = h\left[f_0 + \frac{1}{2}\Delta f_0\right]$$

since all other differences of higher order are absent.

Replacing $\Delta f_0$ by $f_1 - f_0$, we have

$$\int_{x_0}^{x_1} f(x)dx = \frac{h}{2}[f_0 + f_1] \quad (8.7)$$

This is termed as trapezoidal formula of numerical integration.

This formula can be *geometrically interpreted* as the definite integral of the function $f(x)$ between the limits $x_0$ to $x_1$, as is approximated by the area of the trapezoidal region bounded by the chord joining the points $(x_0, f_0)$ and $(x_1, f_1)$, the $x$-axis and the ordinates at $x = x_0$ and at $x = x_1$. This is represented by the shaded area as shown in the Figure 8.1.

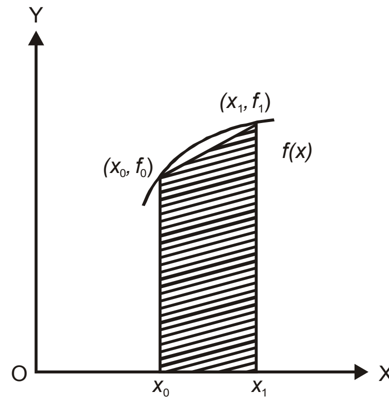



***Fig. 8.1*** *Trapezoidal Region*

Thus, the area under the curve $y = f(x)$ is replaced by the area under the chord joining the points.

The error in the trapezoidal formula is given by,

$$E_T = \frac{h^3}{2} f''(\xi) \times \int_0^1 s(s-1)ds = -\frac{h^3}{12} f''(\xi), \quad \text{where} \quad x_0 < \xi < x_1 \tag{8.8}$$

### Trapezoidal Rule

For evaluating the integral $\int_{x_0}^{x_n} f(x)dx$, we have to sum the integrals for each of the sub-intervals $(x_0, x_1), (x_1, x_2),..., (x_{n-1}, x_n)$. Thus,

$$\int_{x_0}^{x_n} f(x)dx = \frac{h}{2}[(f_0 + f_1) + (f_1 + f_2) + ... + (f_{n-1} + f_n)]$$

Or
$$\int_{x_0}^{x_n} f(x)dx = \frac{h}{2}[f_0 + 2(f_1 + f_2 + ... + f_{n-1}) + f_n] \tag{8.9}$$

This is known as trapezoidal rule of numerical integration.

The error in the trapezoidal rule is,

$$E_T^n = \int_{x_0}^{x_n} f(x)dx - \frac{h}{2}[f_0 + 2(f_1 + f_2 + ... + f_{n-1}) + f_n]$$

$$= \frac{-h^3}{12}[f''(\xi_1) + f''(\xi_2) + ... + f''(\xi_n)]$$

Where $\quad x_0 < \xi_1 < x_1, \quad x_1 < \xi_2 < x_2, ..., \quad x_{n-1} < \xi_n < x_n$

NOTES

Thus, we can write

$$E_T^n = -\frac{h^3}{12}[nf''(\xi)], \quad f''(\xi) \text{ being the mean of } f''(\xi_1), f''(\xi_2), ..., f''(\xi_n)$$

$$= -nh\frac{h^2}{12} \quad f''(\xi)$$

Where $E_T^n = -\frac{h^2}{12} \quad (b-a) \quad f''(\xi), \quad \text{since} \quad nh = b - a$

Or, $x_0 < \xi < x_n$

(8.10)

**Algorithm:** Evaluation of $\int_a^b f(x)dx$ by trapezoidal rule.

**Step 1:** Define function $f(x)$

**Step 2:** Initialize $a, b, n$

**Step 3:** Compute $h = (b–a)/n$

**Step 4:** Set $x = a$, $S = 0$

**Step 5:** Compute $x = x + h$

**Step 6:** Compute $S = S + f(x)$

**Step 7:** Check if $x < b$, then go to Step 4 else go to the next step

**Step 8:** Compute $I = h\ (S + (f(a) + f(b))/2)$

**Step 9:** Output $I, n$

**Simpson's One-Third Formula**

Taking $n = 2$ in the Newton-Cotes formula in Equation (8.4), we get Simpson's one-third formula of numerical integration given by,

$$\int_{x_0}^{x_2} f(x)dx = \quad h\left[2f_0 + \frac{2^2}{2}\Delta f_0 + \frac{1}{12}(2\times 2^3 - 3\times 2^2)\Delta^2 f_0\right]$$

$$= h\left[2f_0 + 2 \quad (f_1 - f_0) + \frac{1}{3} \quad (f_2 - 2f_1 + f_0)\right] \qquad (8.11)$$

$$\therefore \int_{x_0}^{x_2} f(x)dx = \frac{h}{3}[f_0 + 4f_1 + f_2]$$

This is known as Simpson's one-third formula of numerical integration.

The error in Simpson's one-third formula is defined as,

$$E_S = \int_{x_0}^{x_2} f(x)\ dx - \frac{h}{3}(f_0 + 4f_1 + f_2)$$

Assuming $F'(x) = f(x)$, we obtain:

$$E_S = F(x_2) - F(x_0) - \frac{h}{3}(f_0 + 4f_1 + f_2)$$

Expanding $F(x_2) = F(x_0+2h), f_1 = f(x_0+h)$ and $f_2 = f(x_0+2h)$ in powers of $h$, we have:

$$E_S = 2hF'(x_0) + \frac{(2h)^2}{2!}F''(x) \quad (x_0) + \frac{(2h)^3}{3!}F'''(x_0) + ...$$

$$-\frac{h}{3}\left[f_0 + 4\left(f_0 + hf_0' + \frac{h^2}{2!}f''(0) + ...\right) + f_0 + 2hf_0' + \frac{(2h)^2}{2!}f''(0) + ...\right]$$

$$= 2hf_0 + 2h^2 f_0' + \frac{4}{3}h^3 f''(0) + \frac{2}{3}h^4 f'''(0) + \frac{4}{15}h^5 y_0^{iv} \quad (\xi)$$

$$-\frac{h}{3} \quad [6f_0 + 6hf_0' + 4h^2 f''(0) + 2h^3 f'''(0)...]$$

$$E_S = -\frac{h^5}{90}f^{iv}(\xi), \text{ on simplification, where } x_0 < \xi < x_2 \qquad (8.12)$$

Geometrical interpretation of Simpson's one-third formula is that the integral represented by the area under the curve is approximated by the area under the parabola through the points $(x_0, f_0)$, $(x_1, f_1)$ and $(x_2, f_2)$ shown in Figure 8.2.

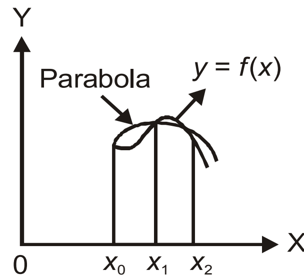



***Fig. 8.2** Simpson's One-Third Integration*

### Simpson's One-Third Rule

On dividing the interval $[a, b]$ into $2m$ sub-intervals by points $x_0 = a, x_1 = a + h$, $x_2 = a + 2h$, ..., $x_{2m} = a+2mh$, where $b = x_{2m}$ and $h = (b-a)/(2m)$, and using Simpson's one-third formula in each pair of consecutive sub-intervals, we have

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + ... + \int_{x_{2m-2}}^{x_{2m}} f(x)dx$$

$$= \frac{h}{3}\left[(f_0 + 4f_1 + f_2) + (f_2 + 4f_3 + f_4) + (f_4 + 4f_5 + f_6) + ... + (f_{2m-2} + 4f_{2m-1} + f_{2m})\right]$$

$$\int_a^b f(x)dx = \frac{h}{3}\left[f_0 + 4(f_1 + f_3 + f_5 + ... + f_{2m-1}) + 2(f_2 + f_4 + f_6 + ... + f_{2m-2}) + f_{2m}\right] \qquad (8.13)$$

NOTES

This is known as Simpson's one-third rule of numerical integration.

The error in this formula is given by the sum of the errors in each pair of intervals as,

$$E_S^{2m} = -\frac{h^5}{90}[f^{iv}(\xi_1) + f^{iv}(\xi_2) + \quad ... + f^{iv}(\xi_m)]$$

Which can be rewritten as,

$$E_S^{2m} = -\frac{h^5}{90}m \quad f^{iv}(\xi), \quad f^{iv}(\xi) \text{ being the mean of } f^{iv}(\xi_1), \quad f^{iv}(\xi_2),..., f^{iv}(\xi_m)$$

Since $2mh = b - a$, we have

$$E_S^{2m} = -\frac{h^4}{180}(b-a)\ f^{iv}(\xi), \text{ where } a < \xi < b. \tag{8.14}$$

**Algorithm:** Evaluation of $\int_a^b f(x)dx$ by Simpson's one-third rule.

**Step 1:** Define $f(x)$

**Step 2:** Input $a, b, n$ (even)

**Step 3:** Compute $h = (b–a)/n$

**Step 4:** Compute $S_1 = f(a) + f(b)$

**Step 5:** Set $S_2 = 0, x = a$

**Step 6:** Compute $x = x + 2h$

**Step 7:** Compute $S_2 = S_2 + f(x)$

**Step 8:** Check If $x < b$ then go to Step 5 else go to next step

**Step 9:** Compute $x = a + h$

**Step 10:** Compute $S_4 = S_4 + f(x)$

**Step 11:** Compute $x = x + 2h$

**Step 12:** Check If $x > b$ go to next Step else go to Step 9

**Step 13:** Compute $I = (S_1 + 4S_4 + 2S_2)h/3$

**Step 14:** Write $I, n$

NOTES

### Simpson's Three-Eighth Formula

Taking $n = 3$, Newton-Cotes formula can be written as,

$$\int_{x_0}^{x_3} f(x)dx = h\int_0^3 (f_0 + u\,\Delta f_0 + \frac{u(u-1)}{2!}\Delta^2 f_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 f_0)du$$

$$= h\left[uf_0 + \frac{u^2}{2}\Delta f_0 + \frac{1}{2}\left(\frac{u^3}{3} - \frac{u^2}{2}\right)\Delta^2 f_0 + \frac{1}{6}\left(\frac{u^4}{4} - u^3 + u^2\right)\Delta^3 f_0\right]_0^3$$

$$= h\left[3y_0 + \frac{9}{2}\Delta y_0 + \frac{9}{4}\Delta^2 y_0 + \frac{3}{8}\Delta^3 y_0\right]$$

$$= h\left[3y_0 + \frac{9}{2}(y_1 - y_0) + \frac{9}{4}(y_2 - 2y_1 + y_0) + \frac{3}{8}(y_3 - 3y_2 + 3y_1 - y_0)\right]$$

$$\int_{x_0}^{x_3} f(x)\ dx = \frac{3h}{8}(y_0 + 3y_1 + y_3)$$

(8.15)

The truncation error in this formula is $-\frac{3h^5}{80}f^{iv}(\xi), x_0 < \xi < x_3$.

This formula is known as Simpson's three-eighth formula of numerical integration.

As in the case of Simpson's one-third rule, we can write Simpson's three-eighth rule of numerical integration as,

$$\int_a^b f(x)\ dx = \frac{3h}{8}[y_0 + 3y_1 + 3y_2 + 2y_3 + 3y_4 + 3y_5 + 2y_6 + ... + 2y_{3m-3} + 3y_{3m-2} + 3y_{3m-1} + y_{3m}]$$

(8.16)

where $h = (b–a)/(3m)$; for $m = 1, 2,...$

i.e., the interval $(b–a)$ is divided into $3m$ number of sub-intervals.

The rule in Equation (8.16) can be rewritten as,

$$\int_a^b f(x)\ dx = \frac{3h}{8}\ [y_0 + y_{3m} + 3(y_1 + y_2 + y_4 + y_5 + ... + y_{3m-2} + y_{3m-1}) + 2(y_3 + y_6 + ... + y_{3m-3})]$$

(8.17)

The truncation error in Simpson's three-eighth rule is

$$\frac{-3h^4}{240}(b-a)f^{iv}(\xi),\quad x_0 < \xi < x_g m$$

NOTES

### Weddle's Formula

In Newton-Cotes formula with $n = 6$ some minor modifications give the Weddle's formula. Newton-Cotes formula with $n = 6$, gives

$$\int_{x_0}^{x_6} y dx = h\left[6y_0 + 18\Delta y_0 + 27\Delta^2 y_0 + 24y\Delta^3 y_0 + \frac{123}{10}\Delta^5 y_0 + \frac{41}{140}\Delta^6 y_0\right]$$

This formula takes a very simple form if the last term $\frac{41}{140}\Delta^6 y_0$ is replaced by $\frac{42}{140}\Delta^6 y_0 = \frac{3}{10}\Delta^6 y_0$. Then the error in the formula will have an additional term $\frac{1}{140}\Delta^6 y_0$. The above formula then becomes,

$$\int_{x_0}^{x_6} y dx = h\left[6y_0 + 18\Delta y_0 + 27\Delta^2 y_0 + 24\Delta^3 y_0 + \frac{123}{10}\Delta^5 y_0 + \frac{3}{10}\Delta^6 y_0\right]$$

$$\therefore \quad \int_{x_0}^{x_6} y dx = \frac{3h}{10} \quad [y_0 + 5y_1 + y_2 + 6y_3 + y_4 + 5y_5 + y_6]$$

(8.18)

On replacing the differences in terms of $y_i$'s, this formula is known as Weddle's formula.

The error Weddle's formula is $-\frac{1}{140}h^7 \cdot y^{(vi)}(\xi)$ (8.19)

Weddle's rule is a composite Weddle's formula, when the number of sub-intervals is a multiple of 6. One can use a Weddle's rule of numerical integration by sub-dividing the interval $(b - a)$ into $6m$ number of sub-intervals, $m$ being a positive integer. The Weddle's rule is,

$$\int_a^b f(x)dx = \frac{3h}{10}[y_0+5y_1+y_2+6y_3+y_4+5y_5+2y_6+5y_7+y_8+6y_9+y_{10}+5y_{11}+\ldots$$

$$+2y_{6m-6}+5y_{6m-5}+y_{6m-4}+6y_{6m-3}+y_{6m-2}+5y_{6m-1}+y_{6m}] \quad (8.20)$$

where $b–a = 6mh$

$$\text{i.e.,} \int_a^b f(x)\ dx = \frac{3h}{10}[y_0 + y_{6m} + 5\ (y_1 + y_5 + y_7 + y_{11} + \ldots + y_{6m-5} + y_{6m-1}) + y_2 + y_4 + y_8 + y_{10} + \ldots$$

$$+ y_{6m-4} + y_{6m-2} + 6(y_3 + y_9 + \ldots + y_{6m-3}) + 2(y_6 + y_{12} + \ldots + y_{-6})]$$

The error in Weddle's rule is given by $-\frac{1}{840}h^6(b-a)y^{(vi)}(\xi)$

(8.21)

NOTES

**Example 1:** Compute the approximate value of $\int_0^2 x^4 dx$ by taking four sub-intervals and compare it with the exact value.

**Solution:** For four sub-intervals of [0, 2], we have $h = \frac{2}{4} = \frac{1}{2} = 0.6$. We tabulate $f(x) = x^4$.

| $x$ | 0 | 0.5 | 1.0 | 1.5 | 2.0 |
|---|---|---|---|---|---|
| $f(x)$ | 0 | 0.0625 | 1.0 | 5.062 | 16.0 |

By trapezoidal rule, we get

$$\int_0^2 x^4 dx \approx \frac{0.5}{2}[0 + 2 \times (0.0625 + 1.0 + 5.062) + 16.0]$$

$$\approx \frac{1}{4}[12.2690 + 16.0] = \frac{28.2690}{4} = 7.0672$$

By Simpson's one-third rule, we get

$$\int_0^2 x^4 dx = \frac{0 \cdot 5}{3}[0 + 4 \times (0.0625 + 5.062) + 2 \times 1.0 + 16.0]$$

$$= \frac{1}{6}[4 \times 5.135 + 18.0] = \frac{38.5380}{6} = 6.4230$$

Exact value $= \frac{2^5}{5} = \frac{32}{5} = 6.4$

Error in the result by trapezoidal rule = $6.4 - 7.0672 = -0.6672$

Error in the result by Simpson's one third rule = $6.4 - 6.4230 = -0.0230$

**Example 2:** Evaluate the following integral:

$\int_0^1 (4x - 3x^2)dx$ by taking $n = 10$ and using the following rules:

(*i*) Trapezoidal rule and (*ii*) Simpson's one-third rule. Also compare them with the exact value and find the error in each case.

**Solution:** We tabulate $f(x) = 4x - 3x^2$, for $x = 0, 0.1, 0.2, ..., 1.0$.

| $x$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $f(x)$ | 0.0 | 0.37 | 0.68 | 0.93 | 1.12 | 1.25 | 1.32 | 1.33 | 1.28 | 1.17 | 1.0 |

NOTES

(*i*) Using trapezoidal rule, we have

$$\int_0^1 (4x-3x^2)\,dx = \frac{0.1}{2}\left[0+2(0.37+0.68+0.93+1.12+1.25+1.32+1.33+1.28+1.17)+1.0\right]$$

$$= \frac{0.1}{2}\times(18.90+1.0) = 0.995$$

(*ii*) Using Simpson's one-third rule, we have

$$\int_0^1 (4x-3x^2)\,dx = \frac{0.1}{3}[0+4(0.37+0.93+1.25+1.33+1.17)+2(0.68+1.12+1.32+1.28)+1.0]$$

$$= \frac{0.1}{3}[4\times 5.05+2\times 4.40+1.0]$$

$$= \frac{0.1}{3}\times[30.0] = 1.00$$

(*iii*) Exact value $= 1.0$

Error in the result by trapezoidal rule is 0.005 and there is no error in the result by Simpson's one-third rule.

**Example 3:** Evaluate $\int_0^1 e^{-x^2}\,dx,$ using (*i*) Simpson's one-third rule with 10 sub-intervals and (*ii*) Trapezoidal rule.

**Solution:** We tabulate values of $e^{-x^2}$ for the 11 points $x = 0, 0.1, 0.2, 0.3, ....,$ 1.0 as given below.

| $x$ | $e^{-x^2}$ | | |
|---|---|---|---|
| 0.0 | 1.00000 | | |
| 0.1 | | 0.990050 | |
| 0.2 | | | 0.960789 |
| 0.3 | | 0.913931 | |
| 0.4 | | | 0.852144 |
| 0.5 | | 0.778801 | |
| 0.6 | | | 0.697676 |
| 0.7 | | 0.612626 | |
| 0.8 | | | 0.527292 |
| 0.9 | | 0.444854 | |
| 1.0 | 0.367879 | | |
| | 1.367879 | 3.740262 | 3.037901 |

Hence, by Simpson's one-third rule we have,

$$\int_0^1 e^{-x^2}\,dx = \frac{h}{3}\ [f_0 + f_{10} + 4\ (f_1 + f_3 + f_5 + f_7 + f_9) + 2\ (f_2 + f_4 + f_6 + f_8)]$$

$$= \frac{0.1}{3}\ [1.367879 + 4 \times 3.740262 + 2 \times 3.037901]$$

$$= \frac{0.1}{3}\ [1.367879 + 14.961048 + 6.075802]$$

$$= \frac{2.2404729}{3} = 0.7468243 \approx 0.746824$$

Using trapezoidal rule, we get

$$\int_0^1 e^{-x^2}\,dx = \frac{h}{2}[f_0 + f_{10} + 2\ (f_1 + f_2 + ... + f_9)]$$

$$= \frac{0.1}{2}\ [1.367879 + 6.778163]$$

$$= 0.4073021$$

**Example 4:** Compute the integral $I = \int_0^4 (x^3 - 2x^2 + 1)dx$, using Simpson's one-third rule taking $h = 1$ and show that the computed value agrees with the exact value. Give reasons for this.

**Solution:** The values of $f(x) = x^3 - 2x^2 + 1$ are tabulated for $x = 0, 1, 2, 3, 4$ as

| $x$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $f(x)$ | 1 | 0 | 1 | 10 | 33 |

The value of the integral by Simpson's one-third rule is,

$$I = \frac{1}{3}\ [1 + 4 \times 0 + 2 \times 1 + 4 \times 10 + 33] = 25\frac{1}{3}$$

The exact value $= \frac{4^4}{4} - 2 \times \frac{4^3}{3} + 1 \times 4 = 25\frac{1}{3}$

Thus, the computed value by Simpson's one-third rule is equal to the exact value. This is because the error in Simpson's one-third rule contains the fourth order derivative and so this rule gives the exact result when the integrand is a polynomial of degree less than or equal to three.

**Example 5:** Compute $\int_{0.1}^{0.5} e^x dx$ by (*i*) Trapezoidal rule and (*ii*) Simpson's one-third rule and compare the results with the exact value, by taking $h = 0.1$.

**NOTES**

**Solution:** We tabulate the values of $f(x) = e^x$ for $x = 0.1$ to $0.5$ with spacing $h = 0.1$.

| $x$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|
| $f(x) = e^x$ | 1.1052 | 1.2214 | 1.3498 | 1.4918 | 1.6847 |

The value of the integral by trapezoidal rule is,

$$I_T = \frac{0.1}{2}[1.1052 + 2(1.2214 + 1.3498 + 1.4918) + 1.6487]$$
$$= \frac{0.1}{2}[2.7539 + 2 \times 4.0630] = 0.5439$$

The value computed by Simpson's one-third rule is,

$$I_S = \frac{0.1}{3}[1.1052 + 4(1.2214 + 1.4918) + 2 \times 1.3498 + 1.6487]$$
$$= \frac{0.1}{3}[2.7539 + 4 \times 2.7132 + 2.6996] = \frac{0.1}{3}[16.3063] = 0.5435$$

$$\text{Exact value} = e^{0.5} - e^{0.1} = 1.6487 - 1.1052 = 0.5435$$

The trapezoidal rule gives the value of the intergal with an error $-0.0004$ but Simpson's one-third rule gives the exact value.

**Example 6:** Compute $\int_0^1 \frac{dx}{1+x}$ using (*i*) Trapezoidal rule (*ii*) Simpson's one-third rule taking 10 sub-intervals. Hence, find $\log_{e^2}$ and compare it with the exact value up to six decimal places.

**Solution:** We tabulate the values of $f(x) = \frac{1}{1+x}$ for $x = 0, 0.1, 0.2, \ldots, 1.0$ as given below:

| $x$ | $y$ | $f(x) = \frac{1}{1+x}$ | | |
|---|---|---|---|---|
| 0.0 | $y_0$ | 1.000000 | | |
| 0.1 | $y_1$ | | 0.9090909 | |
| 0.2 | $y_2$ | | | 0.8333333 |
| 0.3 | $y_3$ | | 0.1692307 | |
| 0.4 | $y_4$ | | | 0.7142857 |
| 0.5 | $y_5$ | | 0.6666667 | |
| 0.6 | $y_6$ | | | 0.6250000 |
| 0.7 | $y_7$ | | 0.5882352 | |
| 0.8 | $y_8$ | | | 0.5555556 |
| 0.9 | $y_9$ | | 0.5263157 | |
| 1.0 | $y_{10}$ | 0.500000 | | |
| | | 1.500000 | 3.4595391 | 2.7281746 |

**NOTES**

(*i*) Using trapezoidal rule, we have

$$\int_0^1 \frac{dx}{1+x} = \frac{h}{2}[f_0 + f_{10} + 2(f_1 + f_2 + f_3 + f_4 + ... + f_0)]$$

$$= \frac{0.1}{2}[1.500000 + 2 \times (3.4595391 + 2.7281745)]$$

$$= \frac{0.1}{2}[1.500000 + 12.3754272] = 0.6437714.$$

(*ii*) Using Simpson's one-third rule, we get

$$\int_0^1 \frac{dx}{1+x} = \frac{h}{3}[f_0 + f_{10} + 4(f_1 + f_3 + ... + f_9) + 2(f_2 + f_4 + ... + f_8)]$$

$$= \frac{0.1}{3}[1.500000 + 4 \times 3.4595391 + 2 \times 2.7281745]$$

$$= \frac{0.1}{3}[1.5 + 13.838156 + 5.456349] = \frac{0.1}{3} \times 20.794505 = 0.6931501$$

(*iii*) Exact value:

$$\int_0^1 \frac{dx}{1+x} = \log_{e^2} = \frac{0.1}{3}\ [1.500000 + 4 \times 3.4595391 + 2 \times 2.7281745]$$

$$= 0.6931472$$

The trapezoidal rule gives the value of the integral having an error 0.693147 – 0.6437714 = 0.0493758, while the error in the value by Simpson's one-third rule is – 0. 000029.

**Example 7:** Compute $\int_0^{\frac{\pi}{2}} \sqrt{\cos\theta}\, d\theta,$ by (*i*) Simpson's rule and (*ii*) Weddle's formula taking six sub-intervals.

**Solution:** Sub-division of $[0, \frac{\pi}{2}]$ into six sub-intervals will have $h = \frac{\pi}{2} \cdot \frac{1}{6} = 15^\circ = 0 \cdot 26179.$ For applying the integration rules we tabulate $\sqrt{\cos\theta}$ .

| $\theta$ | $0^\circ$ | $15^\circ$ | $30^\circ$ | $45^\circ$ | $60^\circ$ | $75^\circ$ | $90^\circ$ |
|---|---|---|---|---|---|---|---|
| $\sqrt{\cos\theta}$ | 1 | 0.98281 | 0.93061 | 0.84089 | 0.70711 | 0.50874 | 0 |

NOTES

(*i*) The value of the integral by Simpson's one-third rule is given by,

$$I_S = \frac{0.26179}{3}[1+4\times(0.98281+0.84089+0.50874)+2\times(0.093061+0.070711)+0)]$$
$$= \frac{0.26179}{3}[1+4\times 2.33244+2\times 1.63772]$$
$$= \frac{0.26179}{3}\times 13.6052 = 1.18723$$

(*ii*) The value of the integral by Weddle's formula is,

$$I_W = \frac{3}{10}\times 0.26179 \quad [1.05+7.45775+5.04534+0.93061+0.070711]$$
$$= 3\times 0.026179 \quad [14.554411] = 1.143059 \approx 1.14306$$

**Example 8:** Evaluate the integral $\int_0^{\frac{\pi}{2}} \sqrt{1-0.162\sin^2\varphi}\, d\varphi$ by Weddle's formula.

**Solution:** On dividing the interval into six sub-intervals, the length of each sub-interval will be $h = \frac{1}{6}\cdot\frac{\pi}{2} = 0\cdot 26179 = 15^{\circ}$. For computing the integral by Weddle's formula, we tabulate $f(\varphi) = \sqrt{1-0\cdot 162\sin^2\varphi}$.

| $\varphi$ | 0° | 15° | 30° | 45° | 60° | 75° | 90° |
|---|---|---|---|---|---|---|---|
| $f\ (\varphi)$ | 1.0 | 0.99455 | 0.97954 | 0.95864 | 0.93728 | 0.92133 | 0.91542 |

The value of the integral by Weddle's formula is given by,

$$I_W = \frac{3\times 0.26179}{10} \quad [1.0+5 \quad (0.99455+0.92133)+0.97954+6\times 0.95864+0.93728+0.91542]$$
$$= 0.078537\times 19.16348 = 1.50504$$

### Computing an Integral to a Desired Accuracy

For evaluating a definite integral correct to a desired accuracy, one has to make a suitable choice of the value of *h*, the length of sub-interval to be used in the formula. There are two ways of determining *h*, by considering the truncation error in the formula to be used for numerical integration or by successive evaluation of the integral by the technique of interval halving and comparing the results.

### Truncation Error Estimation Method

In the truncation error estimation method, the value of *h* to be used is determined by considering the truncation error in the formula for numerical integration. Let *E* be the error tolerance for the integral to be evaluated. Then *h* is chosen by using the condition,

$$|R| < \varepsilon/2$$

NOTES

As an illustration, consider the evaluation of $\int_1^2 \frac{dx}{x}$ using Simpson's one-third rule accurate up to the third decimal place. We may take $\varepsilon = 10^{-3}$.

If we wish to use Simpson's one-third rule, then the truncation error is $R$,

$$R = \frac{h^4}{180}(2-1)f^{iv}(\xi); \quad 1 < \xi < 2$$

Then $h$ is determined by satisfying the condition,

$$\frac{h^4}{180} | f^{iv}(\xi) | < 0 \cdot 5 \times 10^{-3}$$

For the given problem, $f\,(x) = \frac{1}{x}$, thus $f^{iv}(x) = \frac{2 \times 3 \times 4}{x^5}$. Hence,

$$\max_{[1,2]} \left| f^{iv}(x) \right| = 24$$

Thus, $h^4 \times \frac{1 \times 24}{180} < 0 \cdot 5 \times 10^{-3}$ or $h < 0.102$

But $h$ has to be so chosen such so that the interval [1, 2] is divided into an even number of sub-intervals. Hence we may take $h = 0.1 < 0.102$, for which $n$ = 10, i.e., there will be 10 sub-intervals.

The value of the integral is,

$$\int_1^2 \frac{dx}{x} = \frac{0.1}{3} \left[1.0 + 4 \left(\frac{1}{1.1} + \frac{1}{1.3} + \frac{1}{1.5} + \frac{1}{1.7} + \frac{1}{1.9}\right) + 2\left(\frac{1}{1.2} + \frac{1}{1.4} + \frac{1}{1.6} + \frac{1}{1.8}\right) + \frac{1}{2}\right]$$

$$= \frac{0.1}{3} \quad [1.5 + 4 \times 3.4595 + 2 \times 2.7282]$$

$$= \frac{0.1}{3} \times 2.0749 = 0.06931$$ which agrees with the exact value of $\log_{e^2}$.

## Interval Halving Technique

When the estimation of the truncation error is cumbersome, the method of interval halving is used to compute an integral to the desired accuracy.

In the interval halving technique, an integral is first computed for some moderate value of $h$. Then, it is evaluated again for spacing $\frac{h}{2}$, i.e., with double the number of subdivisions. This requires the evaluation of the integrand at the new points of subdivision only and the previous function values with spacing $h$ are also used. Now the difference between the integral $I_h$ and $I_{\frac{h}{2}}$ is used to check the accuracy of the computed integral. If $\left|I_h - I_{h/2}\right| \le \varepsilon$, where $\varepsilon$ is the permissible error, then

NOTES

$I_{h/2}$ is to be taken as the computed value of the integral to the desired accuracy. If the above accuracy is not achieved, i.e., $|I_h - I_{h/2}| > \varepsilon$, then the computation of the integral is made again with spacing $\frac{h}{4}$ and the accuracy condition is tested again. The equation of $I_{h/4}$ will require the evaluation of the integrand at the new points of sub-division only.

***Notes:***

1. The initial choice of $h$ is sometimes taken as $\sqrt[m]{\varepsilon}$ where $m = 2$ for trapezoidal rule and $m = 4$ for Simpson's one-third rule.
2. The method of interval halving is widely used for computer evaluation since it enables a general choice of $h$ together with a check on the computations.
3. The truncation error $R$ can be estimated by using Runge's principle given by,

$R \approx \frac{1}{3}|I_h - I_{h/2}|$ for trapezoidal rule and $R \approx \frac{1}{15}|I_h - I_{h/2}|$ for Simpson's one-third rule.

**Algorithm:** Evaluation of an integral by Simpson's one-third rule with interval halving.

**Step 1:** Set/initialize $a, b, \varepsilon$

[$a, b$ are limits of integration, $\varepsilon$ is error tolerance]

**Step 2:** Set $h = \frac{b-a}{2}$

**Step 3:** Compute $S_1 = f(a) + f(b)$

**Step 4:** Compute $S_4 = f(a + h)$

**Step 5:** Set $S_2 = 0, I_1 = 0$

**Step 6:** Compute $I_2 = \frac{(S_1 + 4S_4 + S_2) \times h}{3}$

**Step 7:** If $(I_2 - I_1) < \varepsilon$, go to Step 17 else go to the next step

**Step 8:** Set $h = \frac{h}{2}$, $I_1 = I_2$

**Step 9:** Compute $S_2 = S_2 + S_4$

**Step 10:** Set $S_4 = 0$

**Step 11:** Set $x = a + h$

**Step 12:** Compute $S_4 = S_4 + f(x)$

**Step 13:** Set $x = x + h$

**Step 14:** If $x < b$, go to Step 12 else go to the next step

**Step 15:** Compute $I_2 = \frac{(S_1 + 2S_2 + 4S_4) \times h}{3}$

NOTES

**Step 16:** Go to step 7

**Step 17:** Write $I_2$, $h$, $\varepsilon$

**Step 18:** End

**Algorithm:** Evaluation of an integral by trapezoidal rule with interval halving.

**Step 1:** Initialize/set $a$, $b$, $\varepsilon$ [$a$, $b$ are limits of integration, $\varepsilon$ is error tolerance]

**Step 2:** Set $h = b – a$

**Step 3:** Compute $S_1 = \frac{f(a) + f(b)}{2}$

**Step 4:** Compute $I_1 = S_1 \times h$

**Step 5:** Compute $x = a + \frac{h}{2}$

**Step 6:** Compute $I_2 = (S_1 + f(x)) \times h$

**Step 7:** If $|I_2 - I_1| < \varepsilon,$ go to Step 13 else go to the next step

**Step 8:** Set $h = \frac{h}{2}$

**Step 9:** Set $x = a + h$

**Step 10:** Set $I_2 = I_2 + h \times f(x)$

**Step 11:** If $x < b$, go to Step 9 else go to next step

**Step 12:** Go to Step 7

**Step 13:** Write $I_2, \ \ h, \ \ \varepsilon$

**Step 14:** End

### Numerical Evaluation of Double Integrals

We consider the evaluation of a double integral,

$$I = \iint\limits_R f(x, y) dx\, dy \tag{8.22}$$

where $R$ is the rectangular region $a \le x \le b, c \le y \le d$. The double integral can be transformed into a repeated integral in the following form,

$$\int_a^b dx \left[ \int_c^d f(x, y) dy \right] \tag{8.23}$$

NOTES

Writing $F(x) = \int_c^d f(x, y)dy$ considered as a function of $x$, we have (8.24)

$$I = \int_a^b F(x)dx \tag{8.25}$$

Now for numerical integration, we can divide the interval $[a, b]$ into $n$ sub-intervals with spacing $h$ and then use a suitable rule of numerical integration.

**Trapezoidal Rule for Double Integral**

By trapezoidal rule, we can write the integral Equation (8.25) as,

$$\int_a^b F(x)\ dx = \frac{h}{2}[F_0 + F_n + 2(F_1 + F_2 + F_3 + ... + F_{n-1})] \tag{8.26}$$

where $x_0 = a, x_n = b,\ h = \frac{b-a}{n}$ and

$$F_i = F(x_i) = \int_0^1 f(x_i, y)\,dy,\ x_i = a + ih \tag{8.27}$$

for $i = 0, 1, 2,..., n$.

Each $F_i$ can be evaluated by trapezoidal rule. For this, the interval $[c, d]$ may be divided into $m$ sub-intervals each of length $k = \frac{c-d}{m}$. Thus we can write,

$$F_i = \frac{k}{2}[f(x_i, y_0) + f(x_i, y_m) + 2\{f(x_i, y_1) + f(x_i, y_2 + ... + f(x_i, y_{m-1})\}] \tag{8.28}$$

$y_0 = c, y_m = d, y_i = c+ik;\ i = 0, 1,..., m.$

This Equation (8.28) can be written in a compact form,

$$F_i = \frac{k}{2}[f_{i0} + f_{im} + 2(f_{i1} + f_{i2} + ... + f_{im-1})]. \tag{8.29}$$

The relation Equations (8.26) and (8.29) together form the trapezoidal rule for evaluation of double integrals.

**Simpson's One-Third Rule for Double Integrals**

For the evaluation of double integrals we can write Simpson's $\frac{1}{3}$ rule. Thus we have,

$$I = \int_a^b F(x)\,dx = \frac{h}{3}[F_0 + F_n + 2(F_2 + F_4 + ... + F_{n-2}) + 4(F_1 + F_3 + ... + F_{n-1})] \tag{8.30}$$

Where $h = \frac{b-a}{n}$, $n$ is even and

$$F_i = F(x_i) = \int_c^d f(x_i, y)dy, \; x_i = a + ih, \text{ for } i = 0, 1, 2, ..., n \quad (8.31)$$

And, $x_0 = a$ and $x_n = b$

For evaluating $I$, we have to evaluate each of the $(n + 1)$ integrals given in Equation (8.31). For evaluation of $F_i$, we can use Simpson's one-third rule by dividing $[c, d]$ into $m$ sub-intervals. $F_i$ can be written as,

$$F_i = \frac{k}{3}[f(x_i, y_0) + f(x_i, y_m) + 2f(x_i, y_2) + f(x_i, y_4) + ... + f(x_i, y_{m-2}) + 4\{f(x_i, y_1) + f(x_i, y_3) + ... + f(x_i, y_{m-1})\}] \quad (8.32)$$

Equation (8.32) can be written in a compact notation as,

$$F_i = \frac{k}{3}[f_{i0} + f_{im} + 2(f_{i2} + f_{i4} + ... + f_{in-2}) + 4(f_{i1}, f_{i3} + ... + f_{im-1})]$$

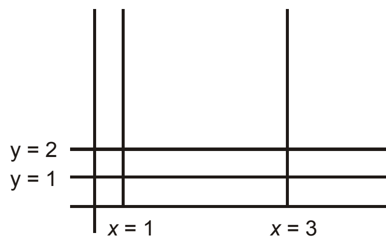
Where $f_{ij} = f(x_i, y_j)$, $j = 0, 1, 2, ..., m$.

**Example 9:** Evaluate the following double integral $\iint_R (x^2 + y^2)dx\,dy$ where $R$ is the rectangular region $1 \leq x \leq 3$, $1 \leq y \leq 2$, by Simpson's one-third rule taking $h = k = 0.5$.

**Solution:** We write the integral in the form of a repeated integral,

$$I = \int_1^3 dx \left[ \int_1^2 (x^2 + y^2)dy \right]$$

Taking $n = 4$ sub-intervals along $x$, so that $h = \frac{2}{4} = 0.5$




$$\therefore I = \int_1^3 F(x)dx = \frac{0.5}{3}[F_0 + F_4 + 2F_2 + 4(F_1 + F_3)]$$

where $F(x) = \int_1^2 (x^2 + y^2)dy$

NOTES

$\therefore \quad F_i = F(x_i) = \int_1^2 (x_i^2 + y^2)dy;$ $x_i = 1 + 0.5i$, where $i = 0, 1, 2, 3, 4.$

For evaluating $F_i$'s, we take $k = \frac{1}{2} = 0.5$ and get,

$$F_0 = \int_1^2 (1+y^2)\,dy = \frac{0.5}{3}[1+1^2+4\{1+(1.5)^2\}+1+2^2] = \frac{0.5}{3}\times 20$$

$$F_1 = \int_1^2 (1.5^2+y^2)\,dy = \frac{0.5}{3}[(1.5)^2+1^2+4\{1.5)^2+(1.5)^2\}+(1.5)^2+2^2] = \frac{0.5}{3}\times 27.50$$

$$F_2 = \int_1^2 (2^2+y^2)\ \ dy = \frac{0.5}{3}\ \ [2^2+1^2+4\ \ (2^2+1.5)^2\}+2^2+2^2] = \frac{0.5}{3}\times 38$$

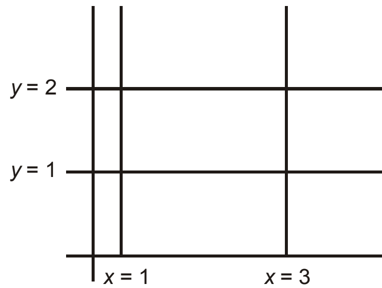
$$F_3 = \int_1^2 ((2.5)^2+y^2)dy = \frac{0.5}{3}\ \ [(2.5)^2+1^2+4\ \ \{(2.5)^2+(1.5)^2\}+(2.5)^2+2^2] = \frac{0.5}{3}\times 51.50$$

$$F_4 = \int_1^2 (3^2+y^2)dy = \frac{0.5}{3}\ \ [3^2+1^2+4\{3^2+(1.5)^2\}+3^2+2^2] = \frac{0.5}{3}\times 68$$

$$\therefore I = \frac{0.25}{9}[20+68+2\times 38+4\ (27.50+51.50)]$$

$$= \frac{0.25}{9}\times 480 = 13.333$$

**Example 10:** Compute $\iint_R (x^2+y^2)dx\,dy$ by trapezoidal rule with $h = 0.5$.




**Solution:** $I_T = \int_1^3 F(x)dx = \frac{0.5}{2}\ [F_0+F_4+2\ (F_1+F_2+F_3)]$

*where* $F_i = F(x_i) = \int_1^2 (x_i^2+y^2)dy,\ x_i = 1 + 0.5i,\ i = 0, 1, 2, 3, 4.$

Thus, $F_0 = \int_1^2 (1+y^2)dy = \frac{0.5}{2}\ \ [1^2+1^2+2\{1^2+(1.5)^2\}+1^2+2^2]$

$$= \frac{0.5}{2}\times 13.50 = 3.375$$

$$F_1 = \int_1^2 [(1.5)^2 + y^2]dy = \frac{0.5}{2} \quad [(1.5)^2 + 1^2 + 2 \quad \{(1.5)^2 + (1.5)^2\} + (1.5)^2 + 2^2]$$

$$= \frac{0.5}{2} \times 18.50 = 4.625$$

$$F_2 = \int_1^2 [2^2 + y^2]dy = \frac{0.5}{2} \quad [2^2 + 1^2 + 2\{2^2 + (1.5)^2\} + 2^2 + 2^2]$$

$$= \frac{0.5}{2} \times 25.50 = 6.375$$

$$F_3 = \int_1^2 [(2.5)^2 + y^2]dy = \frac{0.5}{2} \quad [(2.5)^2 + 1^2 + 2\{(2.5)^2 + (1.5)^2\} + (2.5)^2 + 2^2]$$

$$= \frac{0.5}{2} \times 34.50 = 8.625$$

$$F_4 = \int_1^2 [3^2 + y^2]dy = \frac{0.5}{2} \quad [3^2 + 1^2 + 2\{3^2 + (1.5)^2\} + 3^2 + 2^2]$$

$$= \frac{0.5}{2} \times 45.50 = 11.375$$

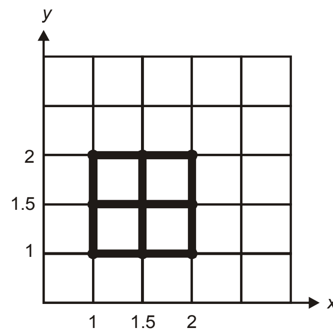
$$\therefore \; I_T = \frac{0.5}{2} \times [3.375 + 11.375 + 2(4.625 + 6.375 + 8.625)]$$

$$= \frac{1}{4}[14.750 + 2 \times 19.625]$$

$$= \frac{1}{4}\left[14.750 + 39.250\right] = \frac{1}{4} \times 54 = 13.5$$

**Example 11:** Evaluate the following double integral using trapezoidal rule with length of sub-intervals $h = k = 0.5, \int_1^2 \int_1^2 \frac{dx \;\, dy}{x+y}.$

**Solution:** Let $f(x, y) = \dfrac{1}{x+y}$



NOTES

By trapezoidal rule with $h = 0.5$, the integral

$$I = \int_1^2 \int_1^2 dx \; dy \; f(x, y) \text{ is computed as,}$$

$$I = \frac{0.5 \times 0.5}{4} \; [f(1, \; 1) + f(2, \; 1) + f(1, \; 2) + f(2, \; 2) + 2\{f(1.5, \; 1) + f(1, \; 1.5)$$

$$+ f(2, \; 1.5) + f(1.5, \; 2)\} + 4f \; (1.5, \; 1.5)]$$

$$= \frac{1}{16}\left[\frac{1}{2} + \frac{1}{3} + \frac{1}{3} + \frac{1}{4} + 2 \; \left(\frac{2}{5} + \frac{2}{5} + \frac{2}{7} + \frac{2}{7}\right) + 4 \times \frac{1}{3}\right]$$

$$= \frac{1}{16}\left[0.666667 + 0.75 + 2 \times \frac{4 \times 12}{35} + \frac{4}{3}\right]$$

$$= \frac{1}{16}[5.492857]$$

$$= 0.343304.$$

**Example 12:** Evaluate $\int_1^2 \int_1^2 \frac{dxdy}{x + y}$ by Simpson's one-third rule. Take sub-intervals of length $h = k = 0.5$.

**Solution:** The value of the integral $I = \int_1^2 \int_1^2 f(x, y) dx \, dy$ by Simpson's one-third rule with $h = k = 0.5$ is,

$$I = \frac{0.5 \times 0.5}{3 \times 3} \; [f(1, \; 1) + f(2, \; 1) + f(1, \; 2) + f(2, \; 2) + 4\{f(1, \; 1.5) + f(1.5, \; 1)$$

$$+ f(2, \; 1.5) + f(1.5, \; 2)\} + 16f(1.5, \; 1.5)]$$

$$= \frac{1}{36} \left[\frac{1}{2} + \frac{1}{3} + \frac{1}{3} + \frac{1}{4} + 4 \; \left(\frac{2}{5} + \frac{2}{5} + \frac{2}{7} + \frac{2}{7}\right) + 16 \times \frac{1}{3}\right]$$

$$= \frac{1}{36}\left[0.666667 + 0.75 + 4 \times \frac{4 \times 12}{35} + \frac{16}{3}\right]$$

$$= \frac{1}{36} \; [12.235714] = 0.339880$$

**Gaussian Quadrature**

We have seen that Newton-Cotes formula of numerical integration is of the form,

$$\int_a^b f(x) dx \approx \sum_{i=0}^{n} c_i f(x_i) \qquad (8.33)$$

where $x_i = a + ih$, $i = 0, 1, 2, ..., n$; $h = \frac{b - a}{n}$

NOTES

This formula uses function values at equally spaced points and gives the exact result for $f(x)$ being a polynomial of degree less than or equal to $n$. Gaussian quadrature formula is similar to Equation (8.33) given by,

$$\int_{-1}^{1} F(u)du \approx \sum_{i=1}^{n} w_i F(u_i) \qquad (8.34)$$

where $w_i$'s and $u_i$'s called weights and abscissae, respectively are derived such that above Equation (8.34) gives the exact result for $F(u)$ being a polynomial of degree less than or equal to $2n–1$.

In Newton-Cotes Equation (8.33), the coefficients $c_i$ and the abscissae $x_i$ are rational numbers but the weights $w_i$ and the abscissae $u_i$ are usually irrational numbers. Even though Gaussian quadrature formula gives the integration of $F(u)$ between the limits –1 to +1, we can use it to find the integral of $f(x)$ from $a$ to $b$ by a simple transformation given by,

$$x = \frac{b-a}{2}u + \frac{a+b}{2} \qquad (8.35)$$

Evidently, then limits for $u$ become –1 to 1 corresponding to $x = a$ to $b$ and writing,

$$f(x) = f\left[\frac{b-a}{2}u + \frac{a+b}{2}\right] = F(u)$$

We have, $\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} F(u)du$

(8.36)

It can be shown that the $u_i$ are the zeros of the Legendre polynomial $P_n(u)$ of degree $n$. These roots are real but irrational and the weights are also irrational.

Given below is a simple formulation of the relevant equations to determine $u_i$ and $w_i$. Let $F(u)$ be a polynomial of the form,

$$F(u) = \sum_{k=0}^{2n-1} a_k u^k \qquad (8.37)$$

Then, we can write

$$\int_{-1}^{1} F(u)du = \int_{-1}^{1}\left[\sum_{k=0}^{2n-1} a_k u^k\right]du \qquad (8.38)$$

Or, $\int_{-1}^{1} F(u)du = 2a_0 + \frac{2}{3}a_2 + \frac{2}{5}a_4 + ... + \frac{2}{2n-2}a_{2n-2}$ (8.39)

NOTES

Equation (8.34) gives,

$$\int_{-1}^{1} F(u)du = \sum_{i=1}^{n} w_i \left[ \sum_{k=0}^{2n-1} a_k u_i^k \right]$$
$$= \sum_{i=1}^{n} w_i \left( a_0 + a_1 u_i + a_2 u_i^2 + ... + a_{2n-1} u_i^{2n-1} \right) \quad (8.40)$$

The Equations (8.39) and (8.40) are assumed to be identical for all polynomials of degree less than or equal to $2n–1$ and hence equating the coefficients of $a_k$ on either side we obtain the following $2n$ equations for the $2n$ unknowns $w_1, w_2,...,w_n$ and $u_1, u_2,...,u_n$.

$$\sum_{i=1}^{n} w_i = 2, \sum_{i=1}^{n} w_i u_i = 0, \sum_{i=1}^{n} w_i u_i^2 = \frac{2}{3}, ... \sum_{i=1}^{n} w_i u_i^{2n-1} = 0 \quad (8.41)$$

The solution of Equation (8.41) is quite complicated. However, use of Legendre polynominals makes the labour unnecessary. It can be shown that the abscissae $u_i$ are the zeros of the Legendre polynomial $P_n(x)$ of degree $n$. The weights $w_i$ can then be easily determined by solving the first $n$ equations of Equations (8.41). As an illustration, we take $n = 2$. The four equations for $u_1$, $u_2$, $w_1$ and $w_2$ are,

$$w_1 + w_2 = 2$$
$$w_1 u_1 + w_2 u_2 = 0$$
$$w_1 u_1^2 + w_2 u_2^2 = \frac{2}{3}$$
$$w_1 u_1^3 + w_2 u_2^3 = 0$$

Eliminating $w_1$, $w_2$, we get

$$\frac{w_1}{w_2} = -\frac{u_2}{u_1} = -\frac{u_2^3}{u_1^3}$$

Or, $u_1^3 u_2 - u_1 u_2^3 = 0$ or $u_1 u_2 (u_1^2 - u_2^2) = 0$

Since, $u_1 \neq u_2 \neq 0$, we have $u_1 = - u_2$.

Also, $w_1 = w_2 = 1$. The third equation gives, $2u_1^2 = \frac{2}{3} \Rightarrow u_1 = \frac{1}{\sqrt{3}}, \ u_2 = -\frac{1}{\sqrt{3}}$

Hence, two point Gauss-Legendre quadrature formula is,

$$\int_{-1}^{1} F(u)du = F\left(\frac{1}{\sqrt{3}}\right) + F\left(-\frac{1}{\sqrt{3}}\right)$$

The Table 8.1 gives the abscissae and weights of the Gauss-Legendre quadrature for values of $n$ from 2 to 6.

NOTES

***Table 8.1*** *Values of Weights and Abscissae for Gauss-Legendre Quadrature*

| $n$ | Weights | Abscissae |
|---|---|---|
| 2 | 1.0 | $\pm 0.57735027$ |
| 3 | 0.88888889 | 0.0 |
| | 0.55555556 | $\pm 0.77459667$ |
| 4 | 0.65214515 | $\pm 0.33998104$ |
| | 0.34785485 | $\pm 0.86113631$ |
| 5 | 0.56888889 | 0.0 |
| | 0.47862867 | $\pm 0.53846931$ |
| | 0.23692689 | $\pm 0.90617985$ |
| 6 | 0.46791393 | $\pm 0.23861919$ |
| | 0.36076157 | $\pm 0.66120939$ |
| | 0.17132449 | $\pm 0.93246951$ |

It is seen that the abscissae are symmetrical with respect to the origin and the weights are equal for equidistant points.

**Example 13:** Compute $\int_0^2 (1+x)dx,$ by Gauss two point quadrature formula.

**Solution:** Substituting $x = u + 1$, the given integral $\int_0^2 (1+x)dx$ reduces to $I = \int_{-1}^{1} (u+2)du$. Using a two point Gauss quadrature formula, we have $I =$ (0.57735027+2) + (– 0.57735027+2) = 4.0.

As expected, the result is equal to the exact value of the integral.

**Example 14:** Show that Gauss two-point quadrature formula for evaluating $\int_a^b f(x)dx$ can be written in the composite form as $\int_a^b f(x)dx = h\sum_{i=0}^{N}[f(r_i)+f(s_i)]$

where $r_i = x_i + hp$, $s_i = x_i + (1-p)h$, $p = \frac{1}{6}(3-\sqrt{3})$.

**Solution:** We subdivide the interval $[a, b]$ into $N$ sub-intervals, each of length $h$, given by $h = \frac{b-a}{N}$.

Consider the integral $I_i$ over the interval $(x_i, x_{i+1})$, i.e., $I_{x_i} = \int_{x_i}^{x_{i+1}} f(x)dx$.

We transform the integral $I_i$ by putting $x = \frac{h}{2}u + \left(x_i + \frac{h}{2}\right)$, so that $x = x_i$ gives $u = -1$ and $x = x_{i+1}$ gives $u = 1$. Thus, $I_i = \frac{h}{2}\int_{-1}^{1} f\left(\frac{h}{2}u + x_i + \frac{h}{2}\right)du$.

NOTES

The Gauss two point quadrature gives,

$$I_i = \frac{h}{2}\left[f\left(\frac{h}{2}\cdot\frac{1}{\sqrt{3}} + x_i + \frac{h}{2}\right) + f\left(-\frac{h}{2\sqrt{3}} + x_i + \frac{h}{2}\right)\right]$$
$$= \frac{h}{2}[f(r_i) + f(s_i)]$$

where $r_i = x_i + ph$, $s_i = x_i + (1 - p)h$, $p = \frac{1}{6}(3-\sqrt{3})$

*Hence,* $\int_a^b f(x)dx = \sum_{i=0}^{N-1} I_i = \frac{h}{2}\sum_{i=0}^{N-1}[f(r_i) + f(s_i)]$

***Note:*** Instead of considering Gauss integration formula for more and more number of points for better accuracy, one can use a two point composite formula for larger number of sub-intervals.

**Example 15:** Evaluate the following integral by Gauss three point quadrature formula:

$$I = \int_0^1 \frac{dx}{1+x}$$

**Solution:** We first transform the interval [0, 1] to the interval (–1, 1) by substituting $t = 2x - 1$, so that $\int_0^1 \frac{dx}{1+x} = \int_{-1}^1 \frac{dt}{t+3}$.

Now by Gauss three point quadrature we have,

$$I = \frac{1}{9}\ [8F(0) + 5F\ (3 + 0.77459667) + 5F\ (3.77459667)] \text{ with } F(t) = \frac{1}{t+3}$$

$\therefore\ I = 0.693122$

The exact value of $\int_0^1 \frac{dx}{1+x} = \ln 2 = 0.693147$

Error = 0.000025

**Romberg's Procedure**

This procedure is used to find a better estimate of an integral using the evaluation of the integral for two values of the width of the sub-intervals.

Let $I_1$ and $I_2$ be the values of an integral $I = \int_a^b f(x)\ dx$, with two different number of sub-intervals of width $h_1$ and $h_2$ respectively using the trapezoidal rule. Let $E_1$ and $E_2$ be the corresponding truncation errors. Since the errors in trapezoidal rule is of order of $h_2$, we can write,

NOTES

$I = I_1 + Kh_1^2$ and $I = I_2 + Kh_2^2$, where $K$ is approximately same.

$$\therefore \quad I_1 + Kh_1^2 = I_2 + Kh_2^2$$

$$\therefore \quad K \approx \frac{I_1 - I_2}{h_2^2 - h_1^2}$$

Thus, $I \approx I_1 + \frac{I_1 - I_2}{h_2^2 - h_1^2}.h_1^2 = \frac{I_1 h_2^2 - I_2 h_1^2}{h_2^2 - h_1^2}$

In Romberg procedure, we take $h_2 = \frac{h_1}{2}$ and we then have,

$$I = \frac{I_1\left(\frac{h_1}{2}\right)^2 - I_2 h_1^2}{\left(\frac{h_1}{2}\right)^2 - h^2} = \frac{4I_2 - I_1}{3}$$

Or,
$$I = I_2 + \left(\frac{I_2 - I_1}{3}\right)$$

This is known as Romberg's formula for trapezoidal integration.

The use of Romberg procedure gives a better estimate of the integral without any more function evaluation. Further, the evaluation of $I_2$ with $h/2$ uses the function values required in evaluation of $I_1$.

**Example 16:** Evaluate $I = \int_0^1 \frac{dx}{1+x^2}$ by trapezoidal rule with $h_1 = 0.5$ and $h_2 = 0.25$ and then use Romberg procedure for a better estimate of $I$. Compare the result with exact value.

**Solution:** We tabulate the value of $x$ and $y = \frac{1}{1+x^2}$ with $h = 0.25$.

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 0 | 0.25 | 0.5 | 0.75 | 1.0 |
| $y$ | 1 | 0.9412 | 0.80 | 0.64 | 0.5 |

Thus using trapezoidal rule, with $h_1 = 0.5$, we have

$$I_1 = \frac{0.5}{3} \times (1 + 0.5 + 2 \times 0.8) = 0.516$$

Similarly, with $h_2 = 0.25$,

$$I_2 = \frac{0.25}{3}[1 + 0.5 + 2 \quad (0.8 + 0.9412 + 0.64)]$$
$$= 0.5218$$

NOTES

The evaluation of $I_2$ uses the function values for evaluation of $I_1$.

By Romberg formula,

$$I \approx I_2 + \frac{1}{3} \ (I_2 - I_1)$$
$$= 0.5218 + (0.5218 - 0.516) \times \frac{1}{3}$$
$$= 0.5218 + 0.0019$$
$$= 0.5237$$

The exact integral $= \tan^{-1} x\Big|_0^1 = \frac{\pi}{4} = 0.5237.$

Thus we can take the result correct to four places of decimals.

**Example 17:** Evaluate $I = \int_1^2 \frac{dx}{x}$ by trapezoidal rule with two and four sub-intervals and then use Romberg procedure to get a better estimate of $I$.

**Solution:** We form a table of value of $y = \frac{1}{x}$ with spacing $h = \frac{1}{4} = 0.25.$

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 1 | 1.25 | 1.5 | 1.75 | 2.0 |
| $y$ | 1 | 0.8 | 0.6667 | 0.5714 | 0.5 |

$$I_1 = \frac{0.5}{2}[1 + 0.5 + 2 \times 0.6667] = 0.7084$$
$$I_2 = \frac{0.25}{2} \ [1 + 0.5 + 2 \ \ (0.8 + 0.6667 + 0.5714)] = 0.6970$$

By Romberg proecedure,

$$I = I_2 + \frac{I_2 - I_1}{3} \approx 0.6970 + \frac{1}{3}(-0.0114)$$
$$= 0.6970 - 0.0038 = 0.6932$$

**Example 18:** Compute the value of $\int_0^1 \frac{dx}{1+x}$, (*i*) by Gauss two point and (*ii*) by Gauss three point formulas.

**Solution:** We first transform the integral by substituting $x = \frac{b-a}{2}t + \frac{1}{2}(b+a)$

$$\int_0^1 \frac{dx}{1+x} = \frac{1}{2}\int_{-1}^1 \frac{1}{1 + \frac{1}{2} + \frac{1}{2}t} = \frac{1}{2}\int_{-1}^1 \frac{2}{3+t}dt$$

NOTES

($i$) By Gauss two point quadrature $\int_{-1}^{1} F(t)\,dt = F\left(\frac{1}{\sqrt{3}}\right) + F\left(-\frac{1}{\sqrt{3}}\right)$ we get,

$$\int_{-1}^{1} \frac{1}{3+t}\,dt = \left(\frac{1}{3+\frac{1}{\sqrt{3}}} + \frac{1}{3-\frac{1}{\sqrt{3}}}\right) = 0.6923$$

($ii$) By Gauss three point quadrature,

$$\int_{-1}^{1} \frac{dt}{3+t}\quad dt = \left[\frac{1}{3}\times 0.888888 + \frac{0.55555556}{3+0.77459667}\right]$$
$$= 0.443478$$

**Example 19:** Compute $\int_{1}^{2} e^x\,dx$ by Gauss three point quadrature.

**Solution:** We first transform the integral by substituting $x = \frac{6-a}{2}t + \frac{1}{2}(b+a) = \frac{1}{2}t + \frac{3}{2}$

$$\therefore \int_{1}^{2} e^x\,dx = \frac{1}{2}\int_{-1}^{1} e^{\frac{t}{2}+\frac{3}{2}}\,dt = \frac{1}{2}e^{\frac{3}{2}}\int_{-1}^{1} e^{\frac{t}{2}}\,dt$$

$$= \frac{1}{2}e^{\frac{3}{2}}\left[0.88888889\times e^0 + 0.55555556\times\left\{e^{\frac{1}{2}}\times 0.77459667 + e^{\frac{1}{2}}\times 0.77459667\right\}\right]$$

$$= 4.67077$$

**Check Your Progress**

1. How will you evaluate a definite integral?
2. Write the trapezoidal formula for numerical integration.
3. What is Simpson's one-third formula of numerical integration?
4. Define Simpson's three-eighth rule of numerical integration.
5. State Weddle's rule.
6. Why is Romberg's procedure used?

## 8.3 NUMERICAL DIFFERENTIATION

Numerical differentiation is the process of computing the derivatives of a function $f(x)$ when the function is not explicitly known, but the values of the function are known only at a given set of arguments $x = x_0, x_1, x_2, \ldots, x_n$. For finding the derivatives, we use a suitable interpolating polynomial and then its derivatives are used as the formulae for the derivatives of the function. Thus, for computing the

derivatives at a point near the beginning of an equally spaced table, Newton's forward difference interpolation formula is used, whereas Newton's backward difference interpolation formula is used for computing the derivatives at a point near the end of the table. Again, for computing the derivatives at a point near the middle of the table, the derivatives of the central difference interpolation formula is used. If, however, the arguments of the table are unequally spaced, the derivatives of the Lagrange's interpolating polynomial are used for computing the derivatives of the function.

### Differentiation Using Newton's Forward Difference Interpolation Formula

Let the values of an unknown function $y = f(x)$ be known for a set of equally spaced values $x_0, x_1, \ldots, x_n$ of $x$, where $x_r = x_0 + r_h$. Newton's forward difference interpolation formula is,

$$\varphi(u) = y_0 + u \ \Delta \ y_0 + \frac{u(u-1)}{2\ !}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3\ !}\Delta^3 y_0 + \ldots + \frac{u(u-1)(u-2)\ldots(u-n+1)}{n\ !}\Delta^n y_0$$

where $u = \dfrac{x - x_0}{h}$

The derivative $\dfrac{dy}{dx}$ can be evaluated as,

$$\frac{dy}{dx} = \frac{d}{dx}\{\varphi(u)\} = \frac{d\varphi}{du} \ . \ \frac{du}{dx} = \frac{1}{h}\frac{d\varphi}{du}$$

Thus, $$y'(x) \simeq \frac{1}{h}\left[\Delta y_0 + \frac{2u-1}{2}\Delta^2 y_0 + \frac{3u^2 - 6u + 2}{6}\Delta^3 y_0 + \frac{2u^3 - 9u^2 + 11u - 3}{12}\Delta^4 y_0 + \ldots\right] \quad (8.42)$$

Similarly, $y''(x) \simeq \dfrac{1}{h^2}\varphi''(u)$

Or, $$y''(x) = \frac{1}{h^2}\left[\Delta^2 y_0 + (u-1)\Delta^3 y_0 + \frac{6u^2 - 18u + 11}{12}\Delta^4 y_0 + \ldots\right] \quad (8.43)$$

For a value of $x$ near the beginning of a table, $u = (x - x_o)/h$ is computed first and then Equation (8.42) and (8.43) can be used to compute $f'(x)$ and $f''(x)$. At the tabulated point $x_0$, the value of $u$ is zero and the formulae for the derivatives are given by,

$$y'(x_0) = \frac{1}{h}\left[\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \frac{1}{5}\Delta^5 y_0 - \ldots\right] \quad (8.44)$$

$$y''(x_0) = \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \frac{5}{6}\Delta^5 y_0 + \ldots\right] \quad (8.45)$$

**Differentiation Using Newton's Backward Difference Interpolation Formula**

For an equally spaced table of a function, Newton's backward difference interpolation formula is,

$$\varphi(v) = y_n + v \ \nabla y_n + \frac{v(v+1)}{2\ !}\nabla^2 y_n + \frac{v(v+1)(v+2)}{3\ !}\nabla^3 y_n + \frac{v(v+1)(v+2)(v+3)}{4\ !}\nabla^4 y_n + ...$$
$$+ \frac{v(v+1)...(v+n-1)}{n\ !}\nabla^n y_n$$

where $v = \frac{x - x_n}{h}$

The derivatives $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$, obtained by differentiating the above formula are given by,

$$\frac{dy}{dx} = \frac{1}{h}\left[\nabla y_n + \frac{2v+1}{2}\nabla^2 y_n + \frac{3v^2+6v+2}{6}\nabla^3 y_n + \frac{2v^3+9v^2+11v+3}{12}\nabla^4 y_n + ...\right] \quad (8.46)$$

$$\frac{d^2y}{dx^2} = \frac{1}{h^2}\left[\nabla^2 y_n + (v+1)\nabla^3 y_n + \frac{6v^2+18v+11}{12}\nabla^4 y_n + ...\right] \quad (8.47)$$

For a given $x$ near the end of the table, the values of $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ are computed by first computing $v = (x - x_n)/h$ and using the above formulae. At the tabulated point $x_n$, the derivatives are given by,

$$y'(x_n) = \frac{1}{h}\left[\nabla y_n + \frac{1}{2}\nabla^2 y_n + \frac{1}{3}\nabla^3 y_n + \frac{1}{4}\nabla^4 y_n + ...\right] \quad (8.48)$$

$$y''(x_n) = \frac{1}{h^2}\left[\nabla^2 y_n + \nabla^3 y_n + \frac{11}{12}\nabla^4 y_n + \frac{5}{6}\nabla^5 y_n + ...\right] \quad (8.49)$$

**Example 20:** Compute the values of $f'(2.1)$, $f''(2.1)$, $f'(2.0)$ and $f''(2.0)$ when $f(x)$ is not known explicitly, but the following table of values is given:

| $x$ | $f(x)$ |
|---|---|
| 2.0 | 0.69315 |
| 2.2 | 0.78846 |
| 2.4 | 0.87547 |

NOTES

**Solution:** Since the points are equally spaced, we form the finite difference table.

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ |
|---|---|---|---|
| 2.0 | 0.69315 | | |
| | | 9531 | |
| 2.2 | 0.78846 | | −83 |
| | | 8701 | |
| 2.4 | 0.87547 | | |

For computing the derivatives at $x = 2.1$, we have

$$f'(x) \simeq \frac{1}{h}[\Delta f_0 + \frac{2u-1}{2}\Delta^2 f_0] \quad \text{and} \quad f''(x) \simeq \frac{1}{h^2}\Delta^2 f_0$$

$$u = \frac{x - x_0}{h} = \frac{2.1 - 2.0}{0.2} = 0.5$$

$$\therefore \quad f'(2.1) = \frac{1}{0.2}\left[0.09531 + \frac{2 \times 0.5 - 1}{2}\Delta^2 f_0\right] = 0.4765$$

$$f''(2.1) = \frac{1}{(0.2)^2} \times (-0.00083) = -0.21$$

The value of $f'(2.0)$ is given by,

$$f'(2.0) = \frac{1}{0.2}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0\right]$$

$$= \frac{1}{0.2}\left[0.09531 + \frac{1}{2} \times 0.00083\right]$$

$$= \frac{0.09572}{0.2} = 0.4786$$

$$f''(2.0) = \frac{1}{(0.2)^2} \times (-0.0083)$$

$$= -0.21$$

**Example 21:** For the function $f(x)$ whose values are given in the table below compute values of $f'(1)$, $f''(1)$, $f'(5.0)$, $f''(5.0)$.

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $f(x)$ | 7.4036 | 7.7815 | 8.1291 | 8.4510 | 8.7506 | 9.0309 |

NOTES

**Solution:** Since $f(x)$ is known at equally spaced points, we form the finite difference table to be used in the differentiation formulae based on Newton's interpolating polynomial.

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ | $\Delta^4 f(x)$ | $\Delta^5 f(x)$ |
|---|---|---|---|---|---|---|
| 1 | 7.4036 | | | | | |
| | | 0.3779 | | | | |
| 2 | 7.7815 | | −303 | | | |
| | | 0.3476 | | 46 | | |
| 3 | 8.1291 | | −257 | | −12 | |
| | | 0.3219 | | 34 | | 8 |
| 4 | 8.4510 | | −223 | | −4 | |
| | | 0.2996 | | 30 | | |
| 5 | 8.7506 | | −193 | | | |
| | | 0.2803 | | | | |
| 6 | 9.0309 | | | | | |

To calculate $f'(1)$ and $f''(1)$, we use the derivative formulae based on Newton's forward difference interpolation at the tabulated point given by,

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \frac{1}{5}\Delta^5 f_0\right]$$

$$f''(x_0) = \frac{1}{h^2}\left[\Delta^2 f_0 - \Delta^3 f_0 + \frac{11}{12}\Delta^4 f_0 - \frac{5}{6}\Delta^5 f_0\right]$$

$$\therefore f'(1) = \frac{1}{1}\left[0.3779 - \frac{1}{2}\times(-0.0303) + \frac{1}{3}\times 0.0046 - \frac{1}{4}\times(-0.0012) + \frac{1}{5}\times 0.0008\right]$$

$$= 0.39507$$

$$f''(1) = \left[0.0303 - 0.0046 + \frac{11}{12}\times(-0.0012) - \frac{5}{6}\times 0.0008\right]$$

$$= -0.0367$$

Similarly, for evaluating $f'(5.0)$ and $f''(5.0)$, we use the following formulae

$$f'(x_n) = \frac{1}{h}\left[\nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{1}{5}\nabla^5 f_n\right]$$

$$f''(x_n) = \frac{1}{h^2}\left[\nabla^2 f_n + \nabla^3 f_n + \frac{11}{12}\nabla^4 f_n + \frac{5}{6}\nabla^5 f_n\right]$$

$$f'(5) = \left[0.2996 + \frac{1}{2}(-0.0223) + \frac{1}{3}\times 0.0034 + \frac{1}{4}(-0.0012)\right]$$

$$= 0.2893$$

$$f''(5) = [-0.0223 + 0.0034 + \frac{11}{12}\times 0.0012]$$

$$= -0.0178$$

NOTES

**Example 22:** Compute the values of $y'(0)$, $y''(0.0)$, $y'(0.02)$ and $y''(0.02)$ for the function $y = f(x)$ given by the following tabular values:

| $x$ | 0.0 | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 |
|---|---|---|---|---|---|---|
| $y$ | 0.00000 | 0.10017 | 0.20134 | 0.30452 | 0.41075 | 0.52110 |

**Solution:** Since the values of $x$ for which the derivatives are to be computed lie near the beginning of the equally spaced table, we use the differentiation formulae based on Newton's forward difference interpolation formula. We first form the finite difference table.

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ |
|---|---|---|---|---|---|
| 0.0 | 0.00000 | | | | |
| | | 0.10017 | | | |
| 0.05 | 0.10017 | | 100 | | |
| | | 0.10117 | | 101 | |
| 0.10 | 0.20134 | | 201 | | 3 |
| | | 0.10318 | | 104 | |
| 0.15 | 0.30452 | | 305 | | 3 |
| | | 0.10623 | | 107 | |
| 0.20 | 0.41075 | | 412 | | |
| | | 0.11035 | | | |
| 0.25 | 0.52110 | | | | |

For evaluating $y'(0,0)$, we use the formula

$$y'(x_0) = \frac{1}{h}\left[\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0\right]$$

$$\therefore \quad y'(0.0) = \frac{1}{0.05}\left(0.10017 - \frac{1}{2}\times 0.00100 + \frac{1}{3}\times 0.00101 - \frac{1}{4}\times 0.00003\right)$$

$$= 2.00000$$

For evaluating $y''(0,0)$, we use the formula

$$y''(x_0) = \frac{1}{h^2}\left(\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0\right)$$

$$= \frac{1}{(0.05)^2}\left(0.00100 - 0.00101 + \frac{11}{12}\times 0.00003\right)$$

$$= 0.007$$

For evaluating $y'(0.02)$ and $y''(0.02)$, we use the following formulae, with

$$u = \frac{0.02 - 0.00}{0.05} = 0.4$$

$$y'(0.02) = \frac{1}{h}\left[\Delta y_0 + \frac{2u-1}{2}\Delta^2 y_0 + \frac{3u^2-6u+2}{6}\Delta^3 y_0 + \frac{2u^3-9u^2+11u-3}{12}\Delta^4 y_0\right]$$

$$y''(0.02) = \frac{1}{h^2}\left[\Delta^2 y_0 + \frac{6(u-1)}{6}\Delta^3 y_0 + \frac{6u^2-18u+11}{12}\times\Delta^4 y_0\right]$$

**NOTES**

$$\therefore \quad y'(0.02) = \frac{1}{0.05}\left[0.10017 + \frac{2\times 0.4-1}{2}\times 0.00100 + \frac{3\times(0.4)^2-6\times 0.4+2}{6}\times 0.00101 + \frac{2\times 0.4^3-9\times 0.4^2+11\times 0.4-3}{12}\times 0.00003\right]$$

$$= 4.00028$$

$$y''(0.02) = \frac{1}{(0.05)^2}\left[0.00100 - 0.00101\times(-0.6) + \frac{6\times 0.16-18\times 0.4+11}{12}\times 0.00003\right]$$

$$= 0.800$$

**Example 23:** Compute $f'(6.0)$ and $f''(6.3)$ by numerical differentiation formulae for the function *f*(*x*) given in the following table.

| $x$ | 6.0 | 6.1 | 6.2 | 6.3 | 6.4 |
|---|---|---|---|---|---|
| $f(x)$ | −0.1750 | −0.1998 | −0.2223 | −0.2422 | −0.2596 |

**Solution:** We first form the finite difference table,

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ |
|---|---|---|---|---|
| 6.0 | −0.1750 | | | |
| | | −248 | | |
| 6.1 | −0.1998 | | 23 | |
| | | −225 | | 3 |
| 6.2 | −0.2223 | | 26 | |
| | | −199 | | −1 |
| 6.3 | −0.2422 | | 25 | |
| | | −174 | | |
| 6.4 | −0.2596 | | | |

For evaluating $f'(6.0)$, we use the formula derived by differentiating Newton's forward difference interpolation formula.

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0\right]$$

$$\therefore \quad f'(6.0) = \frac{1}{0.1}\left[-0.0248 - \frac{1}{2}\times 0.0023 + \frac{1}{3}\times 0.0003\right]$$

$$= 10[-0.0248 - 0.00115 + 0.0001]$$

$$= -0.2585$$

NOTES

For evaluating $f''(6.3)$, we use the formula obtained by differentiating Newton's backward difference interpolation formula. It is given by,

$$f''(x_n) = \frac{1}{h^2}[\nabla^2 f_n + \nabla^3 f_n]$$

$$\therefore \quad f''(6.3) = \frac{1}{(0.1)^2}[0.0026 + 0.0003] = 0.29$$

**Example 24:** Compute the values of $y'(1.00)$ and $y''(1.00)$ using suitable numerical differentiation formulae on the following table of values of $x$ and $y$:

| $x$ | 1.00 | 1.05 | 1.10 | 1.15 | 1.20 |
|---|---|---|---|---|---|
| $y$ | 1.0000 | 1.02470 | 1.04881 | 1.07238 | 1.09544 |

**Solution:** For computing the derivatives, we use the formulae derived on differentiating Newton's forward difference interpolation formula, given by

$$f'(x_0) = \frac{1}{h}\left[\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + ...\right]$$

$$f''(x_0) = \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 + ...\right]$$

Now, we form the finite difference table.

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ |
|---|---|---|---|---|---|
| 1.00 | 1.00000 | | | | |
| | | 2470 | | | |
| 1.05 | 1.02470 | | −59 | | |
| | | 2411 | | 5 | |
| 1.10 | 1.04881 | | −54 | | −2 |
| | | 2357 | | 3 | |
| 1.15 | 1.07238 | | −51 | | |
| | | 2306 | | | |
| 1.20 | 1.09544 | | | | |

Thus with $x_0 = 1.00$, we have

$$y'(1.00) = \frac{1}{0.05}\left(0.02470 + \frac{1}{2}\times 0.00059 + \frac{1}{3}\times 0.00005 + \frac{1}{4}\times 0.00002\right)$$
$$= 0.502$$

$$y''(1.00) = \frac{1}{(0.05)^2}\left(-0.00059 - 0.00005 - \frac{11}{12}\times 0.00002\right)$$
$$= -0.26$$

**Example 25:** Using the following table of values, find a polynomial representation of $f'(x)$ and then compute $f'(0.5)$.

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $f(x)$ | 1 | 3 | 15 | 40 |

**Solution:** Since the values of $x$ are equally spaced we use Newton's forward difference interpolating polynomial for finding $f'(x)$ and $f'(0.5)$. We first form the finite difference table as given below:

| $x$ | $f(x)$ | $\Delta f(x)$ | $\Delta^2 f(x)$ | $\Delta^3 f(x)$ |
|---|---|---|---|---|
| 0 | 1 | | | |
| | | 2 | | |
| 1 | 3 | | 10 | |
| | | 12 | | 3 |
| 2 | 15 | | 13 | |
| | | 25 | | |
| 3 | 40 | | | |

Taking $x_0 = 0$, we have $u = \frac{x - x_0}{h} = x$. Thus the Newton's forward difference interpolation gives,

$$f = f_0 + u\Delta f_0 + \frac{u(u-1)}{2!}\Delta^2 f_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 f_0$$

i.e.,
$$f(x) \approx 1 + 2x + \frac{x(x-1)}{2} \times 10 + \frac{x(x-1)\ \ (x-2)}{6} \times 3$$

or,
$$f(x) = 1 + 3x - \frac{13}{2}x^2 + \frac{1}{2}x^3$$

$\therefore$
$$f'(x) = 3 - 13x + \frac{3}{2}x^2$$

and,
$$f'(0.5) = 3 - 13 \times 0.5 + \frac{3}{2} \times (0.5)^2 = -3.12$$

**Example 26:** The population of a city is given in the following table. Find the rate of growth in population in the year 2001 and in 1995.

| Year $x$ | 1961 | 1971 | 1981 | 1991 | 2001 |
|---|---|---|---|---|---|
| Population $y$ | 40.62 | 60.80 | 79.95 | 103.56 | 132.65 |

**Solution:** Since the rate of growth of the population is $\frac{dy}{dx}$, we have to compute $\frac{dy}{dx}$ at $x = 2001$ and at $x = 1995$. For this we consider the formula for the derivative on approximating $y$ by the Newton's backward difference interpolation given by,

$$\frac{dy}{dx} = \frac{1}{h}\left[\nabla y_n + \frac{2u+1}{2}\nabla^2 y_n + \frac{3u^2+6u+2}{6}\nabla^3 y_n + \frac{2u^3+9u^2+11u+3}{12}\nabla^4 y_n + \ldots\right]$$

Where $u = \frac{x - x_n}{h}$

For this we construct the finite difference table as given below:

NOTES

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ |
|---|---|---|---|---|---|
| 1961 | 40.62 | | | | |
| | | 20.18 | | | |
| 1971 | 60.80 | | −1.03 | | |
| | | 19.15 | | 5.49 | |
| 1981 | 79.95 | | 4.46 | | −4.47 |
| | | 23.61 | | 1.02 | |
| 1991 | 103.56 | | 5.48 | | |
| | | 29.09 | | | |
| 2001 | 132.65 | | | | |

For $x = 2001$, $u = \frac{x - x_n}{h} = 0$

$$\therefore \quad \left(\frac{dy}{dx}\right)_{2001} = \frac{1}{10}\left[29.09 + \frac{1}{2} \times 5.48 + \frac{1}{3} \times 1.02 + \frac{1}{4} \times (-4.47)\right]$$

$$= 3.105$$

For $x = 1995$, $u = \frac{1995 - 1991}{10} = 0.4$

$$\left(\frac{dy}{dx}\right)_{1995} = \frac{1}{10}\left[23.61 + \frac{1.8}{2} \times 4.46 + \frac{3 \times 0.16 + 6 \times 0.4 + 2}{6} \times 5.49\right]$$

$$= 3.21$$

## 8.4 OPTIMUM CHOICE OF STEP LENGTH

In numerical analysis, numerical differentiation describes algorithms for estimating the derivative of a mathematical function or function subroutine using values of the function and perhaps other knowledge about the function. The simplest method is to use finite difference approximations.

An important consideration in practice when the function is calculated using floating-point arithmetic is the choice of step size, $h$. If chosen too small, the subtraction will yield a large rounding error. In fact, all the finite-difference formulae are ill-conditioned and due to cancellation will produce a value of zero if $h$ is small enough. If too large, the calculation of the slope of the secant line will be more accurately calculated, but the estimate of the slope of the tangent by using the secant could be worse.

For the numerical derivative formula evaluated at $x$ and $x + h$, a choice for $h$ that is small without producing a large rounding error is $\sqrt{\varepsilon x}$ (though not when $x = 0$), where the machine epsilon $\varepsilon$ is typically of the order of $2.2 \times 10^{-16}$. A formula for $h$ that balances the rounding error against the secant error for optimum accuracy is,

$$h = 2\sqrt{\varepsilon\left|\frac{f(x)}{f''(x)}\right|}$$

Though not when $f''(x) = 0$, and to employ it will require knowledge of the function.

For single precision the problems are exacerbated because, although $x$ may be a representable floating-point number, $x + h$ almost certainly will not be. This means that $x + h$ will be changed (by rounding or truncation) to a nearby machine-representable number, with the consequence that $(x + h) - x$ will **not equal *h***; the two function evaluations will not be exactly $h$ apart. Consequently, since most decimal fractions are recurring sequences in binary (just as 1/3 is in decimal) a seemingly round step, such as $h = 0.1$ will not be a round number in binary; it is 0.000110011001100...

**Check Your Progress**

7. Define the process of numerical differentiation.
8. Write Newton's forward difference interpolation formula.
9. Write Newton's backward difference interpolation formula.

## 8.5 EXTRAPOLATION METHOD

The interpolating polynomials are usually used for finding values of the tabulated function $y = f(x)$ for a value of $x$ within the table. But, they can also be used in some cases for finding values of $f(x)$ for values of $x$ near to the end points $x_0$ or $x_n$ outside the interval $[x_0, x_n]$. This process of finding values of $f(x)$ at points beyond the interval is termed as extrapolation. We can use Newton's forward difference interpolation for points near the beginning value $x_0$. Similarly, for points near the end value $x_n$, we use Newton's backward difference interpolation formula.

**Example 27:** With the help of appropriate interpolation formula, find from the following data the weight of a baby at the age of one year and of ten years:

| Age = $x$ | 3 | 5 | 7 | 9 |
|---|---|---|---|---|
| Weight = $y$ (kg) | 5 | 8 | 12 | 17 |

NOTES

**Solution:** Since the values of $x$ are equidistant, we form the finite difference table for using Newton's forward difference interpolation formula to compute weight of the baby at the age of required years.

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ |
|---|---|---|---|
| 3 | 5 | | |
| | | 3 | |
| 5 | 8 | | 1 |
| | | 4 | |
| 7 | 12 | | 1 |
| | | 5 | |
| 9 | 17 | | |

Taking $x = 2$, $u = \dfrac{x - x_0}{h} = -0.5$.

Newton's forward difference interpolation gives,

$$y \text{ at } x = 1,\ y(1) = 5 - 0.5 \times 3 + \frac{(-0.5)(-1.5)}{2} \times 1$$
$$= 5 - 1.5 + 0.38 = 3.88 \simeq 3.9 \text{ kg.}$$

Similarly, for computing weight of the baby at the age of ten years, we use Newton's backward difference interpolation given by,

$$v = \frac{x - x_n}{h} = \frac{10 - 9}{2} = 0.5$$
$$y \text{ at } x = 10,\ y(10) = 17 + 0.5 \times 5 + \frac{0.5 \times 1.5}{2} \times 1$$
$$= 17 + 2.5 + 0.38 \simeq 19.88$$

## 8.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The evaluation of a definite integral cannot be carried out when the integrand $f(x)$ is not integrable, as well as when the function is not explicitly known but only the function values are known at a finite number of values of $x$. There are two types of numerical methods for evaluating a definite integral based on the following formula:

$$\int_a^b f(x)\, dx$$

2. The formula is, $\displaystyle\int_{x_0}^{x_1} f(x)dx = \frac{h}{2}[f_0 + f_1]$.

NOTES

3. The formula is, $\int_{x_0}^{x_2} f(x)dx = \frac{h}{3}[f_0 + 4f_1 + f_2]$.

4. Simpson's three-eighth rule of numerical integration is, $\int_a^b f(x)\,dx = \frac{3h}{8}$ $[y_0$ $+ 3y_1 + 3y_2 + 2y_3 + 3y_4 + 3y_5 + 2y_6 + \ldots + 2y_{3m-3} + 3y_{3m-2} + 3y_{3m-1} + y_{3m}]$ where $h = (b–a)/(3m)$; for $m = 1, 2, \ldots$

5. The Weddle's rule is, $\int_a^b f(x)dx = \frac{3h}{10}$ $[y_0 + 5y_1 + y_2 + 6y_3 + y_4 + 5y_5 + 2y_6 +$ $5y_7 + y_8 + 6y_9 + y_{10} + 5y_{11} + \ldots + 2y_{6m-6} + 5y_{6m-5} + y_{6m-4} + 6y_{6m-3} + y_{6m-2} + 5y_{6m-1} + y_{6m}]$, where $b - a = 6mh$.

6. This procedure is used to find a better estimate of an integral using the evaluation of the integral for two values of the width of the sub-intervals.

7. Numerical differentiation is the process of computing the derivatives of a function $f(x)$ when the function is not explicitly known, but the values of the function are known for a given set of arguments $x = x_0, x_1, x_2, \ldots, x_n$. To find the derivatives, we use a suitable interpolating polynomial and then its derivatives are used as the formulae for the derivatives of the function.

8. Newton's forward difference interpolation formula is,

$$\varphi(u) = y_0 + u\,\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + \ldots + \frac{u(u-1)(u-2)\ldots(u-n+1)}{n!}\Delta^n y_0$$

where $u = \frac{x - x_0}{h}$

9. Newton's backward difference interpolation formula is,

$$\varphi(v) = y_n + v\,\nabla y_n + \frac{v(v+1)}{2!}\nabla^2 y_n + \frac{v(v+1)(v+2)}{3!}\nabla^3 y_n + \frac{v(v+1)(v+2)(v+3)}{4!}\nabla^4 y_n + \ldots + \frac{v(v+1)\ldots(v+n-1)}{n!}\nabla^n y_n$$

Where $v = \frac{x - x_n}{h}$

## 8.7 SUMMARY

- Numerical differentiation is the process of computing the derivatives of a function $f(x)$ when the function is not explicitly known, but the values of the function are known only at a given set of arguments $x = x_0, x_1, x_2, \ldots, x_n$.
- For computing the derivatives at a point near the beginning of an equally spaced table, Newton's forward difference interpolation formula is used,

NOTES

whereas Newton's backward difference interpolation formula is used for computing the derivatives at a point near the end of the table.

- Numerical methods can be applied to determine the value of the integral when the integrand is not integrable as well as when the function is not explicitly known but only the function values are known.
- The two types of numerical methods for evaluating a definite integral are Newton-Cotes quadrature and Gaussian quadrature.
- Taking $n = 2$ in the Newton-Cotes formula, we get Simpson's one-third formula of numerical integration while taking $n = 3$, we get Simpson's three-eighth formula of numerical integration.
- In Newton-Cotes formula with $n = 6$ some minor modifications give the Weddle's formula.
- For evaluating a definite integral correct to a desired accuracy, one has to make a suitable choice of the value of $h$, the length of sub-interval to be used in the formula.
- There are two ways of determining $h$, by considering the truncation error in the formula to be used for numerical integration or by successive evaluation of the integral by the technique of interval halving and comparing the results.
- In the truncation error estimation method, the value of $h$ to be used is determined by considering the truncation error in the formula for numerical integration.
- When the estimation of the truncation error is cumbersome, the method of interval halving is used to compute an integral to the desired accuracy.
- Numerical evaluation of double integrals is done by applying trapezoidal rule and Simpson's one-third rule.
- This procedure is used to find a better estimate of an integral using the evaluation of the integral for two values of the width of the sub-intervals.
- For finding the derivatives, we use a suitable interpolating polynomial and then its derivatives are used as the formulae for the derivatives of the function.
- For computing the derivatives at a point near the beginning of an equally spaced table, Newton's forward difference interpolation formula is used, whereas Newton's backward difference interpolation formula is used for computing the derivatives at a point near the end of the table.
- Let the values of an unknown function $y = f(x)$ be known for a set of equally spaced values $x_0, x_1, \ldots, x_n$ of $x$, where $x_r = x_0 + r_h$. Newton's forward difference interpolation formula is,

$$\varphi(u) = y_0 + u\ \Delta\ y_0 + \frac{u(u-1)}{2\ !}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3\ !}\Delta^3 y_0 + \ldots + \frac{u(u-1)(u-2)\ldots(u-n+1)}{n\ !}\Delta^n y_0$$

where $u = \dfrac{x - x_0}{h}$.

**NOTES**

- At the tabulated point $x_0$, the value of $u$ is zero and the formulae for the derivatives are given by,

$$y'(x_0) = \frac{1}{h}\left[\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \frac{1}{5}\Delta^5 y_0 - \ldots\right]$$

$$y''(x_0) = \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \frac{5}{6}\Delta^5 y_0 + \ldots\right]$$

- For a given $x$ near the end of the table, the values of $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ are computed by first computing $v = (x - x_n)/h$ and using the above formulae. At the tabulated point $x_n$, the derivatives are given by,

$$y'(x_n) = \frac{1}{h}\left[\nabla y_n + \frac{1}{2}\nabla^2 y_n + \frac{1}{3}\nabla^3 y_n + \frac{1}{4}\nabla^4 y_n + \ldots\right]$$

$$y''(x_n) = \frac{1}{h^2}\left[\nabla^2 y_n + \nabla^3 y_n + \frac{11}{12}\nabla^4 y_n + \frac{5}{6}\nabla^5 y_n + \ldots\right]$$

- For computing the derivatives at a point near the middle of the table, the derivatives of the central difference interpolation formula is used.
- If the arguments of the table are unequally spaced, then the derivatives of the Lagrange's interpolating polynomial are used for computing the derivatives of the function.

## 8.8 KEY WORDS

- **Newton-Cotes quadrature:** This is based on integrating polynomial interpolation formulae and requires a table of values of the integrand at equally spaced values of the independent variable $x$.
- **Trapezoidal formula:** The trapezoidal formula of numerical integration is defined using the definite integral of the function $f(x)$ between the limits $x_0$ to $x_1$, as it is approximated by the area of the trapezoidal region bounded by the chord joining the points $(x_0, f_0)$ and $(x_1, f_1)$, the $x$-axis and the ordinates at $x = x_0$ and at $x = x_1$.
- **Romberg's procedure:** This procedure is used to find a better estimate of an integral using the evaluation of the integral for two values of the width of the sub-intervals.
- **Weddle's rule:** It is a composite Weddle's formula and is used when the number of sub-intervals is multiple of 6.

NOTES

- **Numerical differentiation:** It is the process of computing the derivatives of a function $f(x)$ when the function is not explicitly known, but the values of the function are known for a given set of arguments $x = x_0, x_1, x_2, ..., x_n$.
- **Newton's forward difference interpolation formula:** The Newton's forward difference interpolation formula is used for computing the derivatives at a point near the beginning of an equally spaced table.
- **Newton's backward difference interpolation formula:** Newton's backward difference interpolation formula is used for computing the derivatives at a point near the end of the table.
- **Central difference interpolation formula:** For computing the derivatives at a point near the middle of the table, the derivatives of the central difference interpolation formula is used.

## 8.9 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. State Newton-Cotes formula.
2. State the trapezoidal rule.
3. What is the difference between Simpson's one-third formula and one-third rule?
4. What is the error in Weddle's rule?
5. Give the truncation error in Simpson's one-third rule.
6. Where is interval halving technique used?
7. Name the methods used for numerical evaluation of double integrals.
8. State the Gauss quadrature formula.
9. State an application of Romberg's procedure.
10. Define the term numerical differentiation.
11. How the derivative $\frac{dy}{dx}$ can be evaluated?
12. Give the formulae for the derivatives at the tabulated point $x0$ where the value of $u$ is zero.
13. Give the differentiation formula for Newton's backward difference interpolation.
14. Give the Newton's backward difference interpolation formula for an equally spaced table of a function.

## Long-Answer Questions

1. Use suitable formulae to compute $y'(1.4)$ and $y''(1.4)$ for the function $y = f(x)$, given by the following tabular values.

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 1.4 | 1.8 | 2.2 | 2.6 | 3.0 |
| $y$ | 0.9854 | 0.9738 | 0.8085 | 0.5155 | 0.1411 |

2. Compute $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ for $x = 1$ where the function $y = f(x)$ is given by the following table:

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
| $y$ | 1 | 8 | 27 | 64 | 125 | 216 |

3. Compute $\int_0^{20} f(x)\, dx$ by Simpson's one-third rule, where:

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 0 | 5 | 10 | 15 | 20 |
| $f(x)$ | 1.0 | 1.6 | 3.8 | 8.2 | 15.4 |

4. Compute $\int_0^{4} x^3\, dx$ by Simpson's one-third formula and comment on the result:

| | | | |
|---|---|---|---|
| $x$ | 0 | 2 | 4 |
| $x^3$ | 0 | 8 | 64 |

5. Compute $\int_0^{2} x^3\, dx$ by Simpson's one-third formula and comment on the result:

6. Compute $\int_0^{2} e^x\, dx$ by Simpson's one-third formula and compare with the exact value, where $e^0 = 1$, $e^1 = 2.72$, $e^2 = 7.39$.

7. Compute an approximate value of $\pi$, by integrating $\int_0^1 \frac{dx}{1+x^2}$, by Simpson's one-third formula.

8. A rod is rotating in a plane about one of its ends. The following table gives the angle $\theta$ (in radians) through which the rod has turned for different values of time $t$ seconds. Find its angular velocity $\frac{d\theta}{dt}$ and angular acceleration $\frac{d^2\theta}{dt^2}$ at $t = 1.0$.

| $t$ secs | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|---|
| $\theta$ radius | 0.0 | 0.12 | 0.48 | 1.10 | 2.00 | 3.20 |

**NOTES**

9. Find $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at $x = 1$ and at $x = 3$ for the function $y = f(x)$, whose values are given in the following table:

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $y$ | 2.7183 | 3.3210 | 4.0552 | 4.9530 | 6.0496 | 7.3891 |

10. Find $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at $x = 0.96$ and at $x = 1.04$ for the function $y = f(x)$ given in the following table:

| $x$ | 0.96 | 0.98 | 1.0 | 1.02 | 1.04 |
|---|---|---|---|---|---|
| $y$ | 0.7825 | 0.7739 | 0.7651 | 0.7563 | 0.7473 |

11. Compute $\int_0^2 (x+1)dx$, by trapezoidal rule by taking four sub-intervals and comment on the result by comparing it with the exact value.

12. Compute $\int_1^{1.4} (x^3+2)dx$, by Simpson's one-third rule by taking four sub-intervals and find the error in the result.

13. Evaluate $\int_0^1 \cos x\, dx$, correct to three significant figures taking five equal sub-intervals.

14. Compute the value of the integral $\int_0^1 \frac{xdx}{1+x}$ correct to three significant figures by Simpson's one-thrid rule with six sub-intervals.

15. Compute the integral $\int_0^1 \frac{dx}{1+x}$, by Simpson's one-third rule taking four sub-intervals and use it to compute the approximate value of $\prod$.

16. Discuss numerical differentiation using Newton's forward difference interpolation formula and Newton's backward difference interpolation formula.

17. Use the following table of values to compute $\int_0^3 f(x)\,dx$:

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $f(x)$ | 1.6 | 3.8 | 8.2 | 15.4 |

NOTES

18. Use suitable formulae to compute $y'(1.4)$ and $y''(1.4)$ for the function $y = f(x)$, given by the following tabular values:

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 1.4 | 1.8 | 2.2 | 2.6 | 3.0 |
| $y$ | 0.9854 | 0.9738 | 0.8085 | 0.5155 | 0.1411 |

19. Compute $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ for $x = 1$ where the function $y = f(x)$ is given by the following table:

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
| $y$ | 1 | 8 | 27 | 64 | 125 | 216 |

20. A rod is rotating in a plane about one of its ends. The following table gives the angle θ (in radians) through which the rod has turned for different values of time $t$ seconds. Find its angular velocity $\frac{d\theta}{dt}$ and angular acceleration $\frac{d^2\theta}{dt^2}$ at $t = 1.0$.

| | | | | | | |
|---|---|---|---|---|---|---|
| $t$ | secs | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| θ | radians | 0.0 | 0.12 | 0.48 | 1.10 | 2.00 | 3.20 |

21. Find $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at $x = 1$ and at $x = 3$ for the function $y = f(x)$, whose values in [1, 6] are given in the following table:

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
| $y$ | 2.7183 | 3.3210 | 4.0552 | 4.9530 | 6.0496 | 7.3891 |

22. Find $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at $x = 0.96$ and at $x = 1.04$ for the function $y = f(x)$ given in the following table:

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 0.96 | 0.98 | 1.0 | 1.02 | 1.04 |
| $y$ | 0.7825 | 0.7739 | 0.7651 | 0.7563 | 0.7473 |

## 8.10 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

NOTES

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

# BLOCK - III
# PDE, ODE AND EULER METHODS

NOTES

# UNIT 9 PARTIAL DIFFERENTIAL EQUATIONS

**Structure**



## 9.0 INTRODUCTION

In this unit, you will learn about partial differential equations. Partial differential equations are used to formulate, and thus aid the solution of, problems involving functions of several variables. Partial differential equations often model multidimensional systems.

You will learn various methods to solve partial differential equations of first, second and higher orders.

## 9.1 OBJECTIVES

After going through this unit, you will be able to:

- Derive partial differential equations of the first order Lagrange's solution
- Know some special type of equations which can be solved easily by methods other than the general method
- Describe Charpit's general method of solution and its special cases

NOTES

- Solve partial differential equations of second and higher orders
- Classify linear partial differential equations of second order
- Explain homogeneous and non-homogeneous equations with constant coefficients
- Reduce partial differential equations to equations with constant coefficients

## 9.2 PARTIAL DIFFERENTIAL EQUATION OF THE FIRST ORDER LAGRANGE'S SOLUTION

**Lagrange's Equation**

The partial differential equation $Pp + Qq = R$, where $P$, $Q$, $R$ are functions of $x$, $y$, $z$, is called Lagrange's linear differential equation.

Form the auxiliary equations $\frac{dx}{P} = \frac{dy}{Q} = \frac{dz}{R}$ and find two indpendent solutions of the auxiliary equations say $u(x, y, z) = C_1$ and $v(x, y, z) = C_2$, where $C_1$ and $C_2$ are constants. Then the solution of the given equation is $F(u, v) = 0$ or $u = F(v)$.

For example, solve $(y^2 + z^2)p - xyq = -xz$

The auxiliary equations are

$$\frac{dx}{y^2 + z^2} = \frac{dy}{-xy} = \frac{dz}{-xz} \tag{9.1}$$

Taking the last two equations, we get

$$\frac{dy}{y} = \frac{dz}{z}$$

Integrating we get $\log y = \log z + \text{constant}$

$$\therefore \quad \frac{y}{z} = C_1$$

Each of the Equations (9.1) is equal to

$$\frac{xdx + ydy + zdz}{x(y^2 + z^2) - xy^2 - xz^2}$$

i.e. $$\frac{xdx + ydy + zdz}{0}$$

i.e. $$xdx + ydy + zdz = 0$$

Hence after integration this reduces to

$$x^2 + y^2 + z^2 = C_2$$

**NOTES**

Hence the general solution of the equation is

$$F\left(\frac{y}{z}, x^2 + y^2 + z^2\right) = 0$$

**Example 1:** Solve $x^2 \dfrac{\partial z}{\partial x} + y^2 \dfrac{\partial z}{\partial y} = (x + y)z$

**Solution:** The auxiliary equations are

$$\frac{dx}{x^2} = \frac{dy}{y^2} = \frac{dz}{(x+y)z}$$

i.e. $$\frac{dx - dy}{x^2 - y^2} = \frac{dz}{(x+y)z}$$

i.e. $$\frac{dx - dy}{x - y} = \frac{dz}{z}$$

i.e. $\log(x-y) = \log z + \text{constant}$

$\therefore$ $$\frac{x-y}{z} = C_1$$

Also $$\frac{dx}{x^2} = \frac{dy}{y^2}$$

Hence $$-\frac{1}{x} = -\frac{1}{y} + \text{constant}$$

$\therefore$ $$\frac{1}{y} - \frac{1}{x} = C_2$$

Hence the solution is, $F\left(\dfrac{1}{y} - \dfrac{1}{x}, \dfrac{x-y}{z}\right) = 0$

**Example 2:** Solve $(x^2 - yz)p + (y^2 - zx)q = z^2 - xy$

**Solution:**

The subsidiary equations are:

$$\frac{dx}{x^2 - yz} = \frac{dy}{y^2 - zx} = \frac{dz}{z^2 - xy}$$

$$\frac{dx - dy}{x^2 - yz - (y^2 - zx)} = \frac{d(x-y)}{(x-y)(x+y+z)}$$

$$= \frac{d(y-z)}{(y-z)(x+y+z)}$$

NOTES

$$\therefore \quad \frac{d(x-y)}{x-y} = \frac{d(y-z)}{y-z}$$

Integrating $\log(x-y) = \log(y-z) + \log C_1$

$$\therefore \quad \frac{x-y}{y-z} = C_1 \quad (1)$$

Using multipliers $x, y, z$, each of the subsidiary equations

$$= \frac{xdx + ydy + zdz}{x^3 + y^3 + z^3 - 3xyz} = \frac{xdx + ydy + zdz}{(x+y+z)(x^2 + y^2 + z^2 - xy - yz - zx)}$$

and is also equal to $\dfrac{dx + dy + dz}{x^2 + y^2 + z^2 - yz - zx - xy}$

$$\therefore \quad \frac{xdx + ydy + zdz}{x+y+z} = \frac{dx + dy + dz}{1}$$

$$xdx + ydy + zdz = (x+y+z)d(x+y+z)$$

On Integrating, we get

$$x^2 + y^2 + z^2 = (x+y+z)^2 + C_2$$

$$\therefore \quad xy + yz + zx = C'_2 \quad (2)$$

From Equations (1) and (2), we get the solution,

$$F\left(\frac{x-y}{y-z}, xy + yz + zx\right) = 0, \text{ where } F \text{ is arbitrary.}$$

**Example 3:** Solve $(a-x)p + (b-y)q = c-z$

**Solution:**

The subsidiary equations are:

$$\frac{dx}{a-x} = \frac{dy}{b-y} = \frac{dz}{c-z} \quad (1)$$

From Equation (1)

$$\frac{dy}{b-y} = \frac{dz}{c-z}$$

i.e. $$\frac{dy}{y-b} = \frac{dz}{z-c}$$

$$\log(y-b) = \log(z-c) + \log C_1$$

$$\therefore \quad \frac{y-b}{z-c} = C_1$$

**NOTES**

Also

$$\frac{dx}{a-x} = \frac{dy}{b-y}$$

$$\therefore \quad \frac{dx}{x-a} = \frac{dy}{y-b}$$

$$\therefore \quad \log(x-a) = \log(y-b) + \log C_2$$

$$\therefore \quad \left(\frac{x-a}{y-b}\right) = C_2$$

The general solution is

$$F\left(\frac{y-b}{z-c}, \frac{x-a}{y-b}\right) = 0$$

**Example 4:** Solve $(y-z)p + (z-x)q = x-y$

**Solution:**

The auxiliary equations are:

$$\frac{dx}{y-z} = \frac{dy}{z-x} = \frac{dz}{x-y} = \frac{dx+dy+dz}{0}$$

$$\therefore \quad dx + dy + dz = 0$$

Integrating we get, $x + y + z = C_1$

Also each ratio

$$= \frac{xdx + ydy + zdz}{x(y-z) + y(z-x) + z(x-y)}$$

$$= \frac{xdx + ydy + zdz}{0}$$

$$\therefore \quad xdx + ydy + zdz = 0$$

On integrating, we get,

$$x^2 + y^2 + z^2 = C_2$$

$\therefore$ The general solution is

$$F(x + y + z, x^2 + y^2 + z^2) = 0$$

**Example 5:** Solve $(mz - ny)p - (nx - lz)q = ly - mx$

**Solution:**

The auxiliary equations are:

$$\frac{dx}{mz-ny} = \frac{dy}{nx-lz} = \frac{dz}{ly-mx}$$

NOTES

Using multipliers $x, y, z$, we get each ratio

$$= \frac{xdx + ydy + zdz}{x(mz - ny) + y(nx - lz) + z(ly - mx)}$$

$$= \frac{xdx + ydy + zdz}{0}$$

$\therefore \quad x^2 + y^2 + z^2 = C_1$

Also by using multipliers $l, m, n$, we get each ratio

$$= \frac{ldx + mdy + ndz}{0}$$

$\therefore \quad lx + my + nz = C_2$

$\therefore$ The general solution is

$$F\left(x^2 + y^2 + z^2, lx + my + nz\right) = 0$$

**Example 6:** Solve $x(y - z)p + y(z - x)q = z(x - y)$

**Solution:**

The auxiliary equations are:

$$\frac{dx}{xy - xz} = \frac{dy}{yz - yx} = \frac{dz}{zx - zy}$$

$$= \frac{dx + dy + dz}{0}$$

$\therefore \quad dx + dy + dz = 0$

On integrating, we get, $x + y + z = C_1$ (1)

$$\frac{\frac{dx}{x}}{y - z} = \frac{\frac{dy}{y}}{z - x} = \frac{\frac{dz}{z}}{x - y} = \frac{\frac{dx}{x} + \frac{dy}{y} + \frac{dz}{z}}{0}$$

$\Rightarrow \quad \frac{dx}{x} + \frac{dy}{y} + \frac{dz}{z} = 0$

On integrating, $\log x + \log y + \log z = \log C_2$

$$xyz = C_2 \qquad (2)$$

From Equations (1) and (2), the general solution is, $F(x + y + z, xyz) = 0$

**Example 7:** Solve $x^2p + y^2q = z^2$

**Solution:**

The auxiliary equations are:

$$\frac{dx}{x^2} = \frac{dy}{y^2} = \frac{dz}{z^2}$$

**NOTES**

$$\therefore \qquad \frac{dx}{x^2} = \frac{dy}{y^2}$$

$$\frac{x^{-1}}{-1} = \frac{y^{-1}}{-1} + C_1$$

$$-\frac{1}{x} = -\frac{1}{y} + C_1$$

$$\frac{1}{y} - \frac{1}{x} = C_1$$

Also

$$\frac{dy}{y^2} = \frac{dz}{z^2}$$

$$\therefore \qquad -\frac{1}{y} = -\frac{1}{z} + C_2$$

$$\frac{1}{z} - \frac{1}{y} = C_2$$

The general solution is

$$F\left(\frac{1}{y} - \frac{1}{x}, \frac{1}{z} - \frac{1}{y}\right) = 0$$

**Example 8:** Solve $(y + z)p + (z + x)q = x + y$

**Solution:**

The auxiliary equations are

$$\frac{dx}{y+z} = \frac{dy}{z+x} = \frac{dz}{x+y}$$

i.e.
$$\frac{dx - dy}{x - y} = \frac{dy - dz}{y - z} = \frac{dz - dx}{z - x}$$

$$= \frac{dx + dy + dz}{2(x + y + z)}$$

Considering first two members and integrating, we get

$$\frac{x - y}{y - z} = C_1$$

Considering first and last members and integrating, we get

$$\log(x - y) = \frac{1}{2}\log(x + y + z) + \log C_2$$

NOTES

$$\log \frac{(x-y)^2}{x+y+z} = \log C_2'$$

$$\frac{(x-y)^2}{x+y+z} = \log C_2'$$

∴ The general solution is

$$F\left(\frac{x-y}{y-z}, \frac{(x-y)^2}{x+y+z}\right) = 0$$

## 9.3 SOLUTION OF SOME SPECIAL TYPES OF EQUATIONS

### Wave Equation

For deriving the equation governing small transverse vibrations of an elastic string, we position the string along the $x$-axis, extend it to its length $L$ and fix it at its ends $x = 0$ and $x = L$. Distort the string and at some instant, say $t = 0$, release it to vibrate. Now the problem is to find the deflection $u(x, t)$ of the string at point $x$ and at any time $t > 0$.

To obtain $u(x, t)$ as the result of a partial differential equation we have to make simplifying assumptions as follows:

1. The string is homogeneous. The mass of the string per unit length is constant. The string is perfectly elastic and hence does not offer any resistance to bending.
2. The tension in the string is constant throughout.
3. The vibrations in the string are small so the slope at each point remains small.

For modeling the differential equation, consider the forces working on a small portion of the string. Let the tension be $T_1$ and $T_2$ at the endpoints $P$ and $Q$ of the chosen portion. The horizontal components of the tension are constant because the points on the string move vertically according to our assumption. Hence we have,

$$T_1 \cos\alpha = T_2 \cos\beta = T = \text{const} \tag{9.2}$$

The two forces in the vertical direction are $-T_1 \sin\alpha$ and $T_2 \sin\beta$ of $T_1$ and $T_2$. The negative sign shows that the component is directed downward. If $\rho$ is the mass of the undeflected string per unit length and $\Delta x$ is the length of that portion of the string that is undeflected then by Newton's second law the resultant of these two forces is equal to the mass $\rho\Delta x$ of the portion times the acceleration $\partial^2 u / \partial t^2$

NOTES

$$T_2 \sin\beta - T_1 \sin\alpha = \rho\Delta x \frac{\partial^2 u}{\partial t^2}.$$

By using Equation (9.1), we can divide the above equation by $T_2 \cos\beta = T_1 \cos\alpha = T$, to get

$$\frac{T_2 \sin\beta}{T_2 \cos\beta} - \frac{T_1 \sin\alpha}{T_1 \cos\alpha} = \tan\beta - \tan\alpha = \frac{\rho\Delta x}{T}\frac{\partial^2 u}{\partial t^2} \tag{9.3}$$

Since $\tan\alpha$ and $\tan\beta$ are the slopes of the string at $x$ and $x + \Delta x$, therefore

$$\tan\alpha = \left.\left(\frac{\partial u}{\partial x}\right)\right|_x \qquad \textit{and} \qquad \tan\beta = \left.\left(\frac{\partial u}{\partial x}\right)\right|_{x+\Delta x}.$$

By dividing Equation (9.3) by $\Delta x$ and substituting the values of $\tan\alpha$ and $\tan\beta$, we have

$$\frac{1}{\Delta x}\left[\left.\left(\frac{\partial u}{\partial x}\right)\right|_{x+\Delta x} - \left.\left(\frac{\partial u}{\partial x}\right)\right|_x\right] = \frac{\rho}{T}\frac{\partial^2 u}{\partial t^2}.$$

As $\Delta x$ approaches zero, the equation becomes the linear partial differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \qquad c^2 = \frac{T}{\rho} \tag{9.4}$$

which is the one-dimensional wave equation governing the vibrations of an elastic string

$$\frac{\partial^2 \mathrm{u}}{\partial \mathrm{t}^2} = \mathrm{c}^2 \frac{\partial^2 \mathrm{u}}{\partial \mathrm{x}^2}, \tag{9.5}$$

To determine the solution we use the boundary conditions, $x = 0$ and $x = L$,

$$u(0,t) = 0, \qquad u(L,t) = 0 \qquad \text{for all } t \tag{9.6}$$

The initial velocity and initial deflection of the string determine the form of motion. If $f(x)$ is the original deflection and $g(x)$ is the initial velocity, then our initial conditions are,

$$u(x,0) = f(x) \tag{9.7}$$

and

$$\left.\frac{\partial u}{\partial t}\right|_{t=0} = g(x). \tag{9.8}$$

NOTES

I. Now the problem is to get the solution of Equation (9.5) satisfying the conditions (9.6)-(9.8).

By using the method of separation of variables, verify solutions of the wave Equation (9.5) of the form

$$u(x,t) = F(x)G(t) \tag{9.9}$$

which are a product of two functions, $F(x)$ and $G(t)$. Note here that each of these functions is dependent on one variable, i.e., either $x$ or $t$. By differentiating Equation (9.9) two times both with respect to $x$ and $t$, we obtain

$$\frac{\partial^2 u}{\partial t^2} = F\ddot{G} \qquad \text{and} \qquad \frac{\partial^2 u}{\partial x^2} = F''G$$

By substituting these values in the wave equation we get,

$$F\ddot{G} = c^2 F''G \cdot$$

Divide this equation by $c^2 FG$, to get

$$\frac{\ddot{G}}{c^2 G} = \frac{F''}{F}.$$

The equations on either side are dependent on different variables. Hence changing $x$ will not change $G$ and changing $t$ will not change $F$ and the other side will remain constant. Thus,

$$\frac{\ddot{G}}{c^2 G} = \frac{F''}{F} = k.$$

or

$$F'' - kF = 0 \tag{9.10}$$

and

$$\ddot{G} - c^2 kG = 0 \cdot \tag{9.11}$$

The constant $k$ is arbitrary.

Now we will find the solutions of Equations (9.10) and (9.11) so that the equation $u = FG$ fulfills the boundary conditions (9.6), that is,

$$u(0,t) = F(0)G(t) = 0,\ u(L,t) = F(L)G(t) = 0 \qquad \text{for all } t.$$

When $G \equiv 0$, then $u \equiv 0$.

Therefore, $G \neq 0$ and

(a) $F(0) = 0$, (b) $F(L) = 0$ (9.12)

For $k = 0$ the general solution of Equation (9.10) is $F = ax + b$, and from Equation (9.12) we obtain $a = b = 0$ and hence $F \equiv 0$, which gives

$u \equiv 0$. But for positive value of $k$, i.e., $k = \mu^2$ the general solution of Equation (9.10) is

$$F = Ae^{\mu,x} + Be^{-\mu x},$$

NOTES

and from Equation (9.12), we again get $F \equiv 0$. Hence choose $k < 0$, i.e., $k = -p^2$. Then the Equation (9.10) becomes,

$$F'' + p^2 F = 0$$

The general solution of the above equation is,

$$F(x) = A\cos px + B\sin px.$$

Using conditions of Equation (9.12), we have

$$F(0) = A = 0 \quad \text{and} \quad F(L) = B\sin pL = 0$$

$B = 0$ implies $F \equiv 0$. Thus we will take $\sin pL = 0$, giving

$$pL = n\pi, \quad \text{so that } p = \frac{n\pi}{L} \text{ where } n \text{ is an integer} \tag{9.13}$$

For $B = 1$, we get infinitely many solutions $F(x) = F_n(x)$, where

$$F_n(x) = \sin\frac{n\pi}{L}x \qquad (\mathrm{n} = 1, 2, \cdots). \tag{9.14}$$

These solutions satisfy Equation (9.12). The value of the constant $k$ is now limited to the values $k - p^2 = -(n\pi/L)^2$, resulting from Equation (9.13), so Equation (9.11) becomes

$$\ddot{G} + \lambda_n^2 G = 0 \quad \textit{where } \lambda_n = \frac{cn\pi}{L}. \tag{9.15}$$

A general solution is

$$G_n(t) = B_n \cos\lambda_n t + B_n{}^* \sin\lambda_n t.$$

Hence solutions of (9.5) satisfying (9.6) are $u_n(x,t) = F_n(x)G_n(t)$, written as

$$u_n(x,t) = (B_n\cos\lambda_n t + B_n{}^*\sin\lambda_n t)\sin\frac{n\pi}{L}x \; (n = 1, 2, \cdots). \tag{9.16}$$

Functions of these type are called the **eigenfunctions** and the values $\lambda_n = cn\pi/L$ are called the eigenvalues of the vibrating string. This set of $\lambda_n$ is known as spectrum.

Each $u_n$ represents a harmonic motion with frequency $\lambda_n / 2\pi = cn/2L$ cycles per unit time. This motion is known as the $n$th normal mode of the string.

NOTES

The first normal mode is referred as the fundamental mode ($n = 1$) while the others are known as overtones.

A single solution $u_n(x,t)$ will not satisfy the initial conditions (9.7) and (9.8). But, $u_n$ is a solution of Equation (9.5), since the equation is linear and homogeneous. To obtain a solution that satisfies Equations (9.7) and (9.8), consider the following infinite series,

$$u(x,t) = \sum_{n=1}^{\infty} u_n(x,t) = \sum_{n=1}^{\infty} (B_n \cos \lambda_n t + B_n * \sin \lambda_n t) \sin \frac{n\pi}{L} x,$$

where $\lambda_n = cn\pi / L$ (9.17)

Therefore,

$$u(x,0) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi}{L} x = f(x). \quad (9.18)$$

Select the coefficients $B_n$'s so that $u(x,0)$ becomes the Fourier sine series of $f(x)$. Thus, from Equation (9.10),

$$B_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx, \; n = 1,2,\cdots. \quad (9.19)$$

Similarly, by differentiating Equation (9.17) with respect to $t$ and using Equation (9.8), we get

$$\left.\frac{\partial u}{\partial t}\right|_{t=0} = \left[ \sum_{n=1}^{\infty} (-B_n \lambda_n \sin \lambda_n t + B_n * \lambda_n \cos \lambda_n t) \sin \frac{n\pi x}{L} \right]_{t=0}$$

$$= \sum_{n=1}^{\infty} B_n * \lambda_n \sin \frac{n\pi x}{L} = g(x)$$

$B_n$ *'s should be selected so that for $t = 0$ the partial derivative $\partial u / \partial t$ becomes the Fourier sine series of the function $g(x)$. So from Equation (9.10),

$$B_n * \lambda_n = \frac{2}{L} \int_0^L g(x) \sin \frac{n\pi x}{L} dr$$

Here, since $\lambda_n = cn\pi / L$,

$$B_n * = \frac{2}{cn\pi} \int_0^L g(x) \sin \frac{n\pi x}{L} dx \;\; n = 1,2,\cdots. \quad (9.20)$$

Now, let us consider the case when the initial velocity $g(x)$ is zero. Then the $B_n$ * are zero and Equation (9.17) becomes,

$$u(x,t)=\sum_{n=1}^{\infty}B_n\cos\lambda_n t\sin\frac{n\pi x}{L},\qquad \lambda_n=\frac{cn\pi}{L}. \tag{9.21}$$

NOTES

We know that,

$$\cos\frac{cn\pi}{L}t\sin\frac{n\pi}{L}x=\frac{1}{2}\left[\sin\left\{\frac{n\pi}{L}(x-ct)\right\}+\sin\left\{\frac{n\pi}{L}(x+ct)\right\}\right]$$

Therefore Equation (9.21) becomes,

$$u(x,t)=\frac{1}{2}\sum_{n=1}^{\infty}B_n\sin\left\{\frac{n\pi}{L}(x-ct)\right\}+\frac{1}{2}\sum_{n=1}^{\infty}B_n\sin\left\{\frac{n\pi}{L}(x-ct)\right\}$$

The above two series are generated by substituting $x - ct$ and $x + ct$, respectively, for the variable $x$ in the Fourier sine series given in Equation (9.18) for $f(x)$. Thus

$$u(x,t)=\frac{1}{2}\left[f^*(x-ct)+f^*(x+ct)\right] \tag{9.22}$$

where $f^*$ is the odd periodic extension of $f$ with the period $2L$. By differentiating Equation (9.22) we see that $u(x, t)$ is a solution of Equation (9.5), given that $f(x)$ is twice differentiable on the interval $0 < f(x) < L$ and has one-sided second derivatives at $x = 0$ and $x = L$, which are zero. $u(x, t)$ is obtained as a solution satisfying Equations (9.6) – (9.8).

If $f'(x)$ and $f''(x)$ are merely piecewise continuous or if the one-sided derivatives are not zero, then for each $t$ there will be finitely many values of $x$ at which the second derivatives of $u$ appearing in Equation (9.5) do not exist. Except at these points the wave equation will still be satisfied. We can then regard $u(x, t)$ as a generalized solution.

**Example 9:** Determine the solution of the wave Equation (9.5) corresponding to the following triangular initial deflection,

$$f(x)=\begin{cases}\dfrac{2k}{L}x & \text{if}\quad 0<x<\dfrac{L}{2}\\[2mm] \dfrac{2k}{L}(L-x) & \text{if}\quad \dfrac{L}{2}<x<L\end{cases}$$

and zero initial velocity.

**Solution:** Since $g(x)\equiv 0$, we have $B_n^*=0$ in Equation (9.17).

The $B_n$ are given by Equation (9.11) and thus Equation (9.17) takes the form $u(x,t)=\dfrac{8k}{\pi^2}\left[\dfrac{1}{1^2}\sin\dfrac{\pi}{L}x\cos\dfrac{\pi c}{L}t-\dfrac{1}{3^2}\sin\dfrac{3\pi}{L}x\cos\dfrac{3\pi c}{L}t+\cdots\right].$

## 9.4 CHARPIT'S GENERAL METHOD OF SOLUTION AND ITS SPECIAL CASES

NOTES

Charpit's method is used to find the solution of most general partial differential equation of order one, given by

$$F(x, y, z, p, q) = 0 \tag{9.23}$$

The primary idea in this method is the introduction of a second partial differential equation of order one,

$$f(x, y, z, p, q, a) = 0 \tag{9.24}$$

containing an arbitrary constant '$a$' and satisfying the following conditions :

1. Equations (9.23) and (9.24) can be solved to give

$$p = p(x, y, z, a) \text{ and } q = q(x, y, z, a)$$

2. The equation

$$dz = p(x, y, z, a)dx + q(x, y, z, a)dy \tag{9.25}$$

is integrable.

When a function '$f$' satisfying the conditions 1 and 2 has been found, the solution of Equation (9.25) containing two arbitrary constants (including '$a$') will be a solution of Equation (9.23). The condition 1 will hold if

$$J = \frac{\partial(F, f)}{\partial(p, q)} = \begin{vmatrix} \frac{\partial F}{\partial p} & \frac{\partial f}{\partial p} \\ \frac{\partial F}{\partial q} & \frac{\partial f}{\partial q} \end{vmatrix} \neq 0 \tag{9.26}$$

Condition 2 will hold when

$$p\left(\frac{\partial p}{\partial z}\right) + q\left(-\frac{\partial p}{\partial z}\right) - \left(\frac{\partial p}{\partial y} - \frac{\partial q}{\partial x}\right) = 0$$

$$\Rightarrow \quad p\frac{\partial q}{\partial z} + \frac{\partial q}{\partial x} = q\frac{\partial p}{\partial z} + \frac{\partial p}{\partial y} \tag{9.27}$$

Substituting the values of $p$ and $q$ as functions of $x$, $y$ and $z$ in Equations (9.23) and (9.24) and differentiating with respect to $x$

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial p}\frac{\partial p}{\partial x} + \frac{\partial F}{\partial q}\frac{\partial q}{\partial x} = 0$$

and $$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial p}\frac{\partial p}{\partial x} + \frac{\partial f}{\partial q}\frac{\partial q}{\partial x} = 0$$

NOTES

Therefore,

$$\left(\frac{\partial F}{\partial p}\frac{\partial f}{\partial q}-\frac{\partial F}{\partial q}\frac{\partial f}{\partial p}\right)\frac{\partial q}{\partial x}=\frac{\partial F}{\partial x}\frac{\partial f}{\partial p}-\frac{\partial F}{\partial p}\frac{\partial f}{\partial x}$$

or $$\frac{\partial q}{\partial x}=\frac{1}{J}\left\{\frac{\partial F}{\partial x}\frac{\partial f}{\partial p}-\frac{\partial F}{\partial p}\frac{\partial f}{\partial x}\right\}$$

Similarly $$\frac{\partial p}{\partial y}=\frac{1}{J}\left\{-\frac{\partial F}{\partial y}\frac{\partial f}{\partial q}+\frac{\partial F}{\partial q}\frac{\partial f}{\partial y}\right\}$$

$$\frac{\partial p}{\partial z}=\frac{1}{J}\left\{-\frac{\partial F}{\partial z}\frac{\partial f}{\partial q}+\frac{\partial F}{\partial q}\frac{\partial f}{\partial z}\right\}$$

and $$\frac{\partial q}{\partial z}=\frac{1}{J}\left\{\frac{\partial F}{\partial z}\frac{\partial f}{\partial p}-\frac{\partial F}{\partial p}\frac{\partial f}{\partial z}\right\} \qquad (9.28)$$

Substituting the values from Equation (9.28) in Equation (9.27)

$$\frac{1}{J}\left[p\left(\frac{\partial F}{\partial z}\frac{\partial f}{\partial p}-\frac{\partial F}{\partial p}\frac{\partial f}{\partial z}\right)+\left(\frac{\partial F}{\partial x}\frac{\partial f}{\partial p}-\frac{\partial F}{\partial p}\frac{\partial f}{\partial x}\right)\right]$$

$$=\frac{1}{J}\left[q\left(-\frac{\partial F}{\partial z}\frac{\partial f}{\partial q}+\frac{\partial F}{\partial p}\frac{\partial f}{\partial z}\right)+\left(-\frac{\partial F}{\partial y}\frac{\partial f}{\partial q}+\frac{\partial F}{\partial q}\frac{\partial f}{\partial y}\right)\right]$$

or $$\left(-\frac{\partial F}{\partial p}\right)\frac{\partial f}{\partial x}+\left(-\frac{\partial F}{\partial q}\right)\frac{\partial f}{\partial y}+\left(-p\frac{\partial F}{\partial p}-q\frac{\partial F}{\partial q}\right)\frac{\partial f}{\partial z}$$

$$+\left(p\frac{\partial F}{\partial z}+\frac{\partial F}{\partial x}\right)\frac{\partial f}{\partial p}+\left(q\frac{\partial F}{\partial z}+\frac{\partial F}{\partial y}\right)\frac{\partial f}{\partial q}=0 \qquad (9.29)$$

The Equation (9.29) being linear in variable $x$, $v$, $z$, $p$, $q$ and $f$ has the following subsidiary equations:

$$\frac{dx}{-\frac{\partial F}{\partial p}}=\frac{dy}{-\frac{\partial F}{\partial q}}=\frac{dz}{-p\frac{\partial F}{\partial p}-q\frac{\partial F}{\partial q}}=\frac{dp}{\frac{\partial F}{dx}+p\frac{\partial F}{\partial z}}=\frac{dq}{\frac{\partial F}{\partial y}+q\frac{\partial F}{\partial z}} \qquad (9.30)$$

If any of the integrals of Equations (9.30) involve $p$ or $q$ then it is of the form of Equation (9.24).

Then we solve Equations (9.23) and (9.24) for $p$ and $q$ and integrate Equation (9.25).

NOTES

**Example 10:** Get complete integral of the equation,

$$p^2 + q^2 - 2px - 2qy + 2xy = 0 \qquad (1)$$

**Solution:** The subsidiary equations are

$$\frac{dp}{2(y-p)} = \frac{dq}{2(x-q)} = \frac{dx}{-2(p-x)} = \frac{dy}{-2(q-y)} \qquad (2)$$

$$\therefore \quad \frac{dp+dq}{2y+2x-2p-2q} = \frac{dx+dy}{2x+2y-2p-2q}$$

$$\therefore \quad dp + dq = dx + dy$$

Integrating, we get

$$p + q = x + y + a$$

where $a$ is constant

$$\therefore \quad (p-x)+(q-y) = a \qquad (3)$$

Equation (1) can also be written as

$$(p-x)^2 + (q-y)^2 = (x-y)^2$$

Now $\{(p-x)-(q-y)\}^2 + \{(p-x)+(q-y)\}^2$

$$= 2\{(p-x)^2 + (q-y)^2\}$$

$$\therefore \quad (p-x)-(q-y) = \sqrt{2(x-y)^2 - a^2} \qquad (4)$$

Adding Equations (3) and (4),

$$(p-x) = \frac{1}{2}a + \frac{1}{2}\sqrt{2(x-y)^2 - a^2}$$

or $$p = \frac{a}{2} + x + \frac{1}{2}\sqrt{2(x-y)^2 - a^2}$$

Similarly subtracting Equation (4) from Equation (3)

$$q = y + \frac{a}{2} - \frac{1}{2}\sqrt{2(x-y)^2 - a^2}$$

$$\therefore \quad dz = pdx + qdy$$

or

$$dz = \left\{\frac{a}{2} + x + \frac{1}{2}\sqrt{2(x-y)^2 - a^2}\right\}dx + \left\{y + \frac{a}{2} - \frac{1}{2}\sqrt{2(x-y)^2 - a^2}\right\}dy$$

$$= \frac{1}{2} d\left(x^2 + y^2\right) + \frac{a}{2} d(x+y) + \frac{1}{2}\sqrt{2(x-y)^2 - a^2}\, d(x-y)$$

On integrating

NOTES

$$z + b = \frac{x^2 + y^2}{2} + \frac{a}{2}(x+y) + \frac{1}{2}\int \left(2U^2 - a^2\right)^{\frac{1}{2}} dU$$

where $U = x - y$ and $b$ is an arbitrary constant

$$z + b = \frac{x^2 + y^2}{2} + \frac{a}{2}\left( x + y + \frac{1}{\sqrt{2}}\left\{ \frac{U\sqrt{U^2 - \frac{a^2}{2}}}{2} - \frac{a^2}{4}\log\left( U + \sqrt{U^2 - \frac{a^2}{2}} \right) \right\} \right)$$

$$= \frac{x^2 + y^2}{2} + \frac{a}{2}(x+y) + \frac{(x-y)\sqrt{2(x-y)^2 - a^2}}{4}$$

$$- \frac{a^2}{4\sqrt{2}}\log\left( (x-y) + \sqrt{(x-y)^2 - \frac{a^2}{2}} \right)$$

**Example 11**: Determine the complete integral of the equation

$$p^2 + q^2 - 2px - 2qy + 1 = 0 \qquad (1)$$

**Solution:** The subsidiary equations are

$$\frac{dx}{-(2 - 2xp)} = \frac{dy}{-(2q - 2y)} = \frac{dp}{-2p} = \frac{dq}{-2q} \qquad (2)$$

With

$$\frac{dp}{p} = \frac{dq}{q}$$

On integrating, we get

$$p = aq \qquad (3)$$

where '$a$' is an arbitrary constant.

Substituting the value of $p$ from Equation (3) in Equation (1)

$$q^2\left(1 + a^2\right) - 2q(ax + y) + 1 = 0$$

$$\therefore \quad q = (ax + y) + \sqrt{(ax+y)^2 - \left(1 + a^2\right)}$$

$$\therefore \quad dz = pdx + qdy$$

**NOTES**

which gives

$$dz = q(adx + dy)$$

$$= d(ax+y)\left\{(ax+y)+\sqrt{(ax+y)^2-(1+a^2)}\right\}$$

Integrating

$$z + b = \frac{1}{2}(ax+y)^2 + \frac{(ax+y)\sqrt{(ax+y)^2-(1+a^2)}}{2}$$

$$-\frac{(a^2+1)}{2}\log\left\{(ax+y)+\sqrt{(ax+y)^2-(1+a^2)}\right\}$$

where $b$ is an arbitrary constant.

**Example 12:** Find Complete Integral of the following equation

$$2(pq + py + qx) + x^2 + y^2 = 0 \qquad (1)$$

**Solution:** The subsidiary equations of Equation (1) are

$$\frac{dx}{-(2q+2y)} = \frac{dy}{-(2p+2x)} = \frac{dp}{(2q+2x)} = \frac{dq}{(2p+2y)} \qquad (2)$$

$\therefore\ dp + dq + dx + dy = 0$

Integrating

$$p + q + x + y = \text{constant} = a \text{ (say)}$$

or $\quad (p + x) + (q + y) = a \qquad (3)$

Equation (1) can be written as

$$2(p+x)(q+y)(x-y)^2 = 0$$

or $\quad (p+x)(q+y) = -\frac{1}{2}(x-y)^2$

$\therefore \quad (p+x)-(q+y) = \sqrt{\{(p+x)+(q+y)\}^2 - 4(p+x)(q+y)}$

$$= \sqrt{a^2 + 2(x-y)^2}$$

Adding Equation (3) and (4),

$$2(p+x) = a + \sqrt{a^2+2(x-y)^2}$$

or $\quad p = -x + \frac{a}{2} + \frac{1}{2}\sqrt{a^2+2(x-y)^2}$

**NOTES**

Subtracting Equation (4) from Equation (3)

$$q = -y + \frac{a}{2} - \frac{1}{2}\sqrt{a^2 + 2(x-y)^2}$$

$\therefore \quad dz = pdx + qdy$

giving

$$dz = -(xdx + ydy) + \frac{a}{2}(dx + dy) + \frac{1}{2}\sqrt{a^2 + 2(x-y)^2}\,d(x-y)$$

$$= -\frac{1}{2}d\left(x^3 + y^3\right) + \frac{a}{2}d(x+y) + \frac{1}{2}\sqrt{a^2 + 2(x-y)^2}\,d(x-y)$$

Integrating the above equation, we get

$$2z + b = -\left(x^2 + y^2\right) + a(x+y) + \sqrt{2}\int \sqrt{\frac{a^2}{2}(x-y)^2}\,d(x-y)$$

$$= -\left(x^2 + y^2\right) + a(x+y) + \frac{\sqrt{2}(x-y)\sqrt{\frac{a^2}{2} + (x-y)^2}}{2}$$

$$+ \sqrt{2}\,\frac{a^2}{4}\log\left\{(x-y) + \sqrt{\frac{a^2}{2} + (x-y)^2}\right\}$$

$$= -\left(x^2 + y^2\right) + a(x+y) + \frac{(x-y)\sqrt{a^2 + 2(x-y)^2}}{2}$$

$$+ \frac{a^2}{2\sqrt{2}}\log\left\{(x-y) + \sqrt{\frac{a^2}{2} + (x-y)^2}\right\}.$$

**Example 13:** Find Complete Integral of the equation,

$$p^2 + q^2 - 2pq\tan h\,2y = \sec h^2\,2y$$

**Solution:** The subsidiary equations are,

$$\frac{dx}{-(2p - 2q\tanh 2y)} = \frac{dy}{(-2q - 2p\tanh 2y)} = \frac{dp}{0}$$

$$= \frac{dq}{-4pq\sec h^2\,2y + 4\sec h^2\,2y\tanh 2y}$$

**NOTES**

$\therefore \quad dp = 0$

or $\quad p = \text{constant} = a$ (say)

Therefore

$$q^2 - 2a \tanh 2y.q + a^2 - \sec h^2 2y = 0$$

$$\therefore \quad q = a \tanh 2y + \sqrt{a^2 \tanh^2 2y - a^2 + \sec h^2 2y}$$

$$= a \tanh 2y + \sqrt{1 - a^2} \sec h2y$$

$\therefore \quad dz = pdz + qdy$

gives

$$dz = adx + \left(a \tanh 2y + \sqrt{1-a^2} \sec h2y\right)dy$$

$$= d\left(ax + \frac{a}{2}\log\cosh 2y\right) + \sqrt{1-a^2} \sec h2ydy$$

Integrating

$$z + b = ax + \frac{a}{2}\log\cosh 2y + \sqrt{1-a^2}\int \frac{2dy}{e^{2v} + e^{-2v}}$$

$$= ax + \frac{a}{2}\log\cosh 2y + \sqrt{1-a^2}\int \frac{2e^{2v}dy}{1+e^{4v}}$$

$$= ax + \frac{a}{2}\log\cosh 2y + \sqrt{1-a^2}\left(\tan^{-1} e^{2v}\right).$$

**Example 14:** Find Complete Integral

$$xy + 3yq = 2\left(z - x^2q^2\right) \qquad (1)$$

**Solution:** The subsidiary equations are

$$\frac{dx}{-x} = \frac{dy}{-3y - 4x^3q} = \frac{dp}{p - 2p + 4xq^2} = \frac{dq}{3q - 2q}$$

$$\therefore \quad \frac{dq}{q} = \frac{dx}{-x}$$

$\Rightarrow \quad qx = \text{constant} = a$

$$\Rightarrow \quad q = \frac{a}{x}$$

Substituting in Equation (1) we get

$$p = \frac{2\left(z - a^3\right)}{x} - \frac{3ya}{x^2}$$

NOTES

$\therefore \quad dz = pdx + qdy$

gives $dz = \left\{\frac{2(z-a^2)}{x} - \frac{3ya}{x^2}\right\}dx + \frac{a}{x}dy$

Multiplying by $x^2$

$$x^3 dz = 2x(z-a^2)dx - 3yadx + axdy$$

i.e., $x^4 d\left(\frac{z-a^2}{x^2}\right) = -3aydx + axdy$

i.e., $d\left(\frac{z-a^2}{x^2}\right) = \frac{a}{x^3}dy - \frac{3ay}{x^4}dx = d\left(\frac{ay}{x^2}\right)$

On integrating , we get $\frac{z-a^2}{x^2} = \frac{ay}{x^3} + b$

or $z = a\left(a + \frac{y}{x}\right) + bx^2$ where, $a$ and $b$ are arbitrary constants.

**Check Your Progress**

1. Define Lagrange's linear differential equation.
2. What are the assumptions for solving the wave equation?
3. What is the $n$th normal mode of the string?
4. Where is Charpit's method used?

## 9.5 PARTIAL DIFFERENTIAL EQUATIONS OF SECOND AND HIGHER ORDERS

The general form of a linear differential equation of $n$th order is

$$\frac{d^n y}{dx^n} + P_1 \frac{d^{n-1} y}{dx^{n-1}} + P_2 \frac{d^{n-2} y}{dx^{n-2}} + ... + P_{n-1}\frac{dy}{dx} + P_n y = Q$$

where $P_1, P_2 ..., P_n$ and $Q$ are functions of $x$ alone or constants.

The linear differential equation with constant coefficients are of the form

$$\frac{d^n y}{dx^n} + P_1 \frac{d^{n-1} y}{dx^{n-1}} + P_2 \frac{d^{n-2} y}{dx^{n-2}} + ... + P_{n-1}\frac{dy}{dx} + P_n y = Q \quad (9.31)$$

where $P_1, P_2, ..., P_n$ are constants and $Q$ is a function of $x$.

The equation

$$\frac{d^n y}{dx^n} + P_1 \frac{d^{n-1} y}{dx^{n-1}} + P_2 \frac{d^{n-2} y}{dx^{n-2}} + \ldots + P_{n-1} \frac{dy}{dx} + P_y y = 0 \qquad (9.32)$$

is then called the *Reduced Equation* (R.E.) of the Equation (9.31)

**NOTES**

If $y = y_1(x), y = y_2(x), \ldots, y = y_n(x)$ are $n$-solutions of this reduced equation, then $y = c_1 y_1 + c_2 y_2 + \ldots + c_n y_n$ is also a solution of the reduced equation where $c_1, c_2, \ldots, c_n$ are artbitrary constants.

The solution $y = y_1(x), y = y_2(x), y = y_3(x), \ldots, y = y_n(x)$ are said to be linearly independent if the *Wronskian* of the functions is not zero where the Wronskian of the functions $y_1, y_2, \ldots, y_n$, denoted by $W(y_1, y_2, \ldots, y_n)$, is defined by

$$W(y_1, y_2, \ldots y_n) = \begin{vmatrix} y_1 & y_2 & y_3 \ldots y_n \\ y_1' & y_2' & y_3' \ldots y_n' \\ y_1'' & y_2'' & y_3'' \ldots y_n'' \\ \vdots & \vdots & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & y_3^{(n-1)} \ldots y_n^{(n-1)} \end{vmatrix}$$

Since the general solution of a differential equation of $n$th order contains $n$ arbitrary constants, $u = c_1 y_1 + c_2 y_2 + \ldots + c_n y_n$ is its complete solution.

Let $v$ be any solution of the differential Equation (9.31), then

$$\frac{d^n v}{dx^n} + P_1 \frac{d^{n-1} v}{dx^{n-1}} + P_2 \frac{d^{n-2} v}{dx^{n-2}} + \ldots + P_{n-1} \frac{dv}{dx} + P_n v = Q \qquad (9.33)$$

Since $u$ is a solution of Equation (9.32), we get

$$\frac{d^n u}{dx^n} + P_1 \frac{d^{n-1} u}{dx^{n-1}} + P_2 \frac{d^{n-2} u}{dx^{n-2}} + \ldots + P_{n-1} \frac{du}{dx} + P_n u = 0 \qquad (9.34)$$

Now adding Equation (9.33) and (9.34), we get

$$\frac{d^n (u+v)}{dx^n} + P_1 \frac{d^{n-1}(u+v)}{dx^n} + P_2 \frac{d^{n-2}(u+v)}{dx^{n-2}} + \ldots + P_{n-1} \frac{d(u+v)}{dx} + P_n (u+v) = Q$$

This shows that $y = u + v$ is the complete solution of the Equation (9.31).

Introducing the operators $D$ for $\frac{d}{dx}$, $D^2$ for $\frac{d^2}{dx^2}$, $D^3$ for $\frac{d^3}{dx^3}$ etc. The Equation (9.31) can be written in the form

$$D^n y + P_1 D^{n-1} y + P_2 D^{n-2} y + \ldots\ldots + P_{n-1} Dy + y P_n = Q$$

or $\quad (D^n + P_1 D^{n-1} + P_2 D^{n-2} + \ldots\ldots + P_{n-1} D + P_n) y = Q$

or $\quad F(D) y = Q$ where $F(D) = D^n + P_1 D^{n-1} P_2 D^{n-2} + \ldots\ldots + P_{n-1} D + P_n$

From the above discussions it is clear that the general solution of $F(D)y = Q$ consists of two parts:

(*i*) The Complementary Function (C.F.) which is the complete primitive of the Reduced Equation (R.E.) and is of the form

$y = c_1 y_1 + c_2 y_2 + \ldots + c_n y_n$ containing $n$ arbitrary constants.

(*ii*) The Particular Integral (P.I.) which is a solution of $F(D)\, y = Q$ containing no arbitrary constant.

NOTES

**Rules for Finding The Complementary Function**

Let us consider the 2nd order linear differential equation

$$\frac{d^2 y}{dx^2} + P_1 \frac{dy}{dx} + P_2 y = 0 \qquad (9.35)$$

Let $y = A\, e^{mx}$ be a trial solution of the Equation (9.35); then the auxiliary Equation (A.E.) of Equation (9.35) is given by

$$m^2 + P_1 m + P_2 = 0 \qquad (9.36)$$

The Equation (9.36) has two roots $m = m_1$, $m = m_2$. We discuss the following cases:

(*i*) When $m_1 \neq m_2$, then the complementary function will be

$y = c_1 e^{m_1 x} + c_2\, e^{m_2 x}$ where $c_1$ and $c_2$ are arbitrary constants.

(*ii*) When $m_1 = m_2$, then the complementary function will be

$y = (c_1 + c_2\, x)\, e^{m_1 x}$ where $c_1$ and $c_2$ are arbitrary constants.

(*iii*) When the auxiliary Equation (9.36) has complex roots of the form $\alpha + i\beta$ and $\alpha - i\beta$, then the complementary function will be

$$y = e^{\alpha x}\,(c_1 \cos \beta\, x + c_2 \sin \beta\, x)$$

Let us consider the equation of order $n$

$$\frac{d^n y}{dx^n} + P_1 \frac{d^{n-1} y}{dx^{n-1}} + P_2 \frac{d^{n-2} y}{dx^{n-2}} + \ldots + P_{n-1} \frac{dy}{dx} + P_n y = 0 \qquad (9.37)$$

Let $y = A\, e^{mx}$ be a trial solution of Equation (9.37), then the auxiliary equation is

$$m^n + P_1\, m^{n-1} + P_2\, m^{n-2} + \ldots\ldots + P_{n-1}\, m + P_n = 0 \qquad (9.38)$$

**Rule (1):** If $m_1, m_2, m_3, \ldots, m_n$ be $n$ distinct real roots of Equation (9.38), then the general solution will be

$$y = c_1\, e^{m_1 x} + c_2 e^{m_2 x} + c_3 e^{m_3 x} + \ldots + c_n e^{mnx}$$

where $c_1, c_2, c_3 \ldots\ldots c_n$ are arbitrary constants.

**Rule (2):** If the two roots $m_1$ and $m_2$ of the auxiliary equation are equal and each equal to $m$, the corresponding part of the general solution will be $(c_1 + c_2\, x)\, e^{mx}$ and if the three roots $m_3, m_4, m_5$ are equal to $\alpha$ the corresponding part of the solution is $(c_3 + c_4 x + c_5 x^2)\, e^{\alpha x}$ and others are distinct, the general solution will be

$$y = (c_1 + c_2 x)\, e^{mx} + (c_3 + c_4 + c_5 x^2)\, e^{\alpha x} + c_6\, e^{m_6 x} + \ldots\ldots + c_n e^{mnx}$$

**Rule (3):** If a pair of imaginary roots $\alpha \pm i\beta$ occur twices, the corresponding part of the general solution will be

$$e^{\alpha x}\, [(c_1 + c_2 x) \cos\ \beta x + (c_3 + c_4 x) \sin \beta x]$$

NOTES

and the general solution will be

$$y = e^{\alpha x}[(c_1 + c_2x)\cos\beta x + (c_3 + c_4x)\sin\beta x] + c_5e^{m_5x} + ......+ c_ne^{m_nx}$$

where $c_1, c_2..., c_n$ are arbitrary constants and $m_5, m_6, ...., m_n$ are distinct real roots of (9.38).

**Rule (4):** If the two roots (real) be $m$ and $-m$, the corresponding part of the general solution will be $c_1e^{mx} + c_2e^{-mx}$

$$= c_1(\cosh mx + \sinh mx) + c_2(\cosh mx - \sinh mx)$$

$$= c_1'\cosh mx + c_2'\sinh mx \text{ where } c_1' = c_1 + c_2, c_2' = c_1 - c_2$$

and general solution will be

$$y = c_1'\cosh mx + c_2'\sinh mx + c_3e^{m_3x} + c_4e^{m_4x} + ......+ c_ne^{m_nx}$$

where $c_1', c_2', c_3, .....c_n$ are arbitrary constants and $m_3, m_4 ... m_n$ are distinct real roots of Equation (9.38).

**Rules for Finding Particular Integrals**

Any particular solution of $F(D)y = f(x)$ is known as its Particular Integral (P.I). The P.I. of $F(D)y = f(x)$ is symbolically written as

$$\text{P.I.} = \frac{1}{F(D)}\{f(x)\} \text{ where } F(D) \text{ is the operator.}$$

The operator $\frac{1}{F(D)}$ is defined as that operator which, when operated on $f(x)$ gives a function $\phi(x)$ such that $F(D)\phi(x) = f(x)$

i.e., $$\frac{1}{F(D)}\{f(x)\} = \phi(x) \ (= \text{P.I.})$$

$$\therefore \quad F(D)\left\{\frac{1}{F(D)}f(x)\right\} = f(x) \quad \left[\because \frac{1}{F(D)}f(x) = \phi(x)\right]$$

Obviously $F(D)$ and $1/F(D)$ are inverse operators.

**Case I:** Let $F(D) = D$, then $\frac{1}{D}f(x) = \int f(x)\,dx$.

**Proof:** Let $y = \frac{1}{D}\{f(x)\}$, operating by $D$, we get $Dy = D\,.\frac{1}{D}\{f(x)\}$ or $Dy = f(x)$ or

$$\frac{dy}{dx} = f(x) \text{ or } dy = f(x)\,dx$$

Integrating both sides with respect to $x$, we get

$y = \int f(x)\,dx$, since particular integrating does not contain any arbitrary constant.

**Case II:** Let $F(D) = D - m$ where $m$ is a constant, then

$$\frac{1}{D-m}\{f(x)\} = e^{mx}\int e^{-mx}f(x)dx.$$

NOTES

**Proof:** Let $\frac{1}{D-m}\{f(x)\} = y$, then operating by $D-m$, we get

$$(D-m) \cdot \frac{1}{D-m}\{f(x)\} = (D-m)\, y$$

or $$f(x) = \frac{dy}{dx} - my$$

or $\frac{dy}{dx} - my = f(x)$ which is a first order linear differential equation and I.F. = $e^{\int -mdx} = e^{-mx}$.

Then multiplying above equation by $e^{-mx}$ and integrating with respect to $x$, we get

$y\, e^{-mx} = \int f(x) e^{-mx} dx$, since particular integral does not contain any arbitrary constant

or $$y = e^{mx} \int f(x) e^{-mx} dx.$$

**Note:** If $\frac{1}{F(D)} = \frac{a_1}{D-m_1} + \frac{a_2}{D-m_2} + ..... + \frac{a_n}{D-m_n}$ where $a_i$ and $m_i$ $(i = 1, 2, ..., n)$ are constants, then

$$\frac{1}{F(D)}\{f(x)\} = a_1 e^{m_1 x} \int f(x) e^{-m_1 x} dx + a_2 e^{m_2 x} \int f(x) e^{-m_2 x} dx + ... + a_n e^{m_n x} \int f(x) e^{-m_n x} dx$$

$$= \sum_{i=1}^{n} a_i e^{m_i x} \int f(x) e^{-m_i x} dx$$

We now discuss methods of finding particular integrals for certain specific types of right hand functions

**Type I:** $f(D)\, y = e^{mx}$ where $m$ is a constant.

Then $$\text{P.I.} = \frac{1}{F(D)}\{e^{mx}\} = \frac{e^{mx}}{F(m)} \text{ if } F(m) \neq 0$$

If $F(m) = 0$, then we replace $D$ by $D + m$ in $F(D)$,

$$\text{P.I.} = \frac{1}{F(D)}\{e^{mx}\} = e^{mx} \cdot \frac{1}{F(D+m)}\{1\}$$

**Example 15:** $(D^3 - 2D^2 - 5D + 6)\, y = (e^{2x} + 3)^2 + e^{3x} \cosh x.$

**Solution:** The reduced equation is

$$(D^3 - 2D^2 - 5D + 6)\, y = 0 \qquad ...(1)$$

Let $y = Ae^{mx}$ be a trial solution of (1). Then the auxiliary equation is

$$m^3 - 2m^2 - 5m + 6 = 0 \text{ or } m^3 - m^2 - m^2 + m - 6m + 6 = 0$$

or $\quad m^2(m-1) - m(m-1) - 6(m-1) = 0$

or $\quad (m-1)(m^2 - m - 6) = 0$ or $(m-1)(m^2 - 3m + 2m - 6) = 0$

or $\quad (m-1)(m-3)(m+2) = 0$ or $m = 1, 3, -2$

**NOTES**

$\therefore$ The complementary function is

$y = c_1 e^x + c_2 e^{3x} + c_3 e^{-2x}$ where $c_1, c_2, c_3$ are arbitrary constants.

Again $(e^{2x} + 3)^2 + e^{3x} \cosh x = e^{4x} + 6e^{2x} + 9 + e^{3x}\left(\dfrac{e^x + e^{-x}}{2}\right)$.

$$= e^{4x} + 6e^{2x} + 9e^{0.x} + \frac{e^{4x}}{2} + \frac{e^{2x}}{2}$$

$$= \frac{3}{2}e^{4x} + \frac{13}{2}e^{2x} + 9e^{0.x}$$

$\therefore$ The particular integral is

$$y = \frac{1}{D^3 - 2D^2 - 5D + 6}\left\{\frac{3}{2}e^{4x} + \frac{13}{2}e^{2x} + 9e^{0.x}\right\}$$

$$= \frac{1}{(D-1)(D-3)(D+2)}\left\{\frac{3}{2}e^{4x} + \frac{13}{2}e^{2x} + 9e^{0.x}\right\}$$

$$= \frac{3}{2}\frac{1}{(D-1)(D-3)(D+2)}e^{4x} + \frac{13}{2}\frac{1}{(D-1)(D+2)(D-3)}\{e^{2x}\}$$

$$+ 9\frac{1}{(D-1)(D-3)(D+2)}e^{0.x}$$

$$= \frac{3}{2}\frac{e^{4x}}{(4-1)(4-3)(4+2)} + \frac{13}{2}\frac{e^{2x}}{(2-1))(2+2)(2-3)}$$

$$+ 9\frac{e^{0.x}}{(0-1)(0-3)(0+2)}$$

$$= \frac{3}{2}\frac{e^{4x}}{3.1.6} + \frac{13}{2}\frac{e^{2x}}{1.4.(-1)} + 9\frac{e^{0.x}}{(-1)(-3).2}$$

$$= \frac{e^{4x}}{12} - \frac{13}{8}e^{2x} + \frac{3}{2}.$$

Hence the general solution is

$$y = \text{C.F.} + \text{P.I.}$$

$$= c_1 e^x + c_2 e^{3x} + c_3 e^{-2x} + \frac{e^{4x}}{12} - \frac{13}{8}e^{2x} + \frac{3}{2}.$$

***Notes***: 1. When $F(m) = 0$ and $F'(m) \neq 0$, P.I. $= \dfrac{1}{F(D)}\{e^{mx}\} = x\dfrac{1}{F'(D)}\{e^{mx}\}$

$$= \frac{xe^{mx}}{F'(m)}$$

NOTES

2. When $F(m) = 0$ $F'(m) = 0$ and $F''(m) \neq 0$, then P.I. $= \frac{1}{F(D)}\{e^{mx}\}$

$$= x^2 \frac{1}{F''(D)}\{e^{mx}\} = \frac{x^2 e^{mx}}{F''(m)}$$

and so on.

**Type II:** $f(x) = e^{mx} V$ where $V$ is any function of $x$.

Here the particular integral (P.I.) of $F(D)\, y = f(x)$ is

$$\text{P.I.} = \frac{1}{F(D)}\{e^{mx}V\} = e^{mx}\frac{1}{F(D+m)}\{V\}.$$

**Example 16:** Solve $(D^2 - 5D + 6)\, y = x^2 e^{3x}$

**Solution:** The reduced equation is

$$(D^2 - 5D + 6)\, y = 0 \qquad (1)$$

Let $y = Ae^{mx}$ be a trial solution of Equation (1) and then auxiliary equation is

$$m^2 - 5m + 6 = 0 \text{ or } m^2 - 3m - 2m + 6 = 0$$

or $\quad m(m-3) - 2(m-3) = 0$ or $(m-3)(m-2) = 0$

$\therefore \quad m = 2, 3$

$\therefore$ The complementary function is

$y = c_1 e^{2x} + c_2 e^{3x}$ where $c_1$ and $c_2$ are arbitrary constants.

The particular integral is

$$y = \frac{1}{D^2 - 5D + 6}\{x^2 e^{3x}\} = \frac{e^{3x}}{(D+3)^2 - 5(D+3) + 6}\{x^2\}$$

$$= e^{3x}\frac{1}{D^2 + 6D + 9 - 5D - 15 + 6}\{x^2\} = e^{3x}\frac{1}{D^2 + D}\{x^2\}$$

$$= e^{3x}\frac{1}{D(1+D)}\{x^2\} = e^{3x}\frac{1}{D}(1+D)^{-1}\{x^2\}$$

$$= \frac{e^{3x}}{D}(1 - D + D^2 - D^3 + D^4 - \ldots)\{x^2\}$$

$$= \frac{e^{3x}}{D}\{x^2 - 2x + 2\} = e^{3x}\left(\frac{x^3}{3} - x^2 + 2x\right)$$

Hence the general solution is

$$y = \text{C.F.} + \text{P.I.}$$

$$= c_1 e^{2x} + c_2 e^{3x} + e^{3x}\left(\frac{x^3}{3} - x^2 + 2x\right).$$

**Recall:** (*i*) $(1+x)^{-1} = 1 - x + x^2 - x^3 + x^4 - x^5 + \ldots$

(*ii*) $(1-x)^{-1} = 1 + x + x^2 + x^3 + x^4 + x^5 + \ldots$

**NOTES**

**Type III:** ($a$) $F(D)\,y = \sin ax$ or $\cos ax$ where $F(D) = \phi(D^2)$.

Here P.I. $= \dfrac{1}{F(D)}\{\sin ax\} = \dfrac{1}{\phi(-a^2)}\sin ax$ (if $\phi(-a^2) \neq 0$)

or P.I. $= \dfrac{1}{F(D)}\{\cos ax\} = \dfrac{1}{\phi(-a^2)}\cos ax$ (if $\phi(-a^2) \neq 0$)

[Note $D^2$ has been replaced by $-a^2$ but $D$ has not been replaced by $-a$.]

($b$) $F(D)\,y = \sin ax$ or $\cos ax$ and $F(D) = \phi(D^2, D)$

Here P.I. $= \dfrac{1}{F(D)}\{\sin ax\} = \dfrac{1}{\phi(D^2, D)}\{\sin ax\} = \dfrac{1}{\phi(-a^2, D)}\{\sin ax\}$

if $\phi(-a^2, D) \neq 0$

or $y = \dfrac{1}{F(D)}\{\cos ax\} = \dfrac{1}{\phi(D^2, D)}\{\cos ax\} = \dfrac{1}{\phi(-a^2, D)}\{\cos ax\}$

if $\phi(-a^2, D) \neq 0$

($c$) $F(D)\,y = \sin ax$ or $\cos ax$ and $F(D) = \dfrac{\psi(D)}{\phi(D^2)}$

Here P.I. $= \dfrac{1}{F(D)}\{\sin ax\} = \dfrac{\psi(D)}{\phi(D^2)}\{\sin ax\} = \dfrac{\psi(D)}{\phi(-a^2)}\{\sin ax\}$ if $\phi(-a^2) \neq 0$

or $y = \dfrac{1}{F(D)}\{\cos ax\} = \dfrac{\psi(D)}{\phi(D^2)}\{\cos ax\}$

$= \dfrac{\psi(D)}{\phi(-a^2)}\{\cos ax\}$ if $\phi(-a^2) \neq 0$

($d$) $F(D)\,y = \sin ax$ or $\cos ax$, $F(D) = \phi(D^2)$ but $\phi(-a^2) = 0$.

Here P.I. $= \dfrac{1}{F(D)}\{\sin ax \text{ or } \cos ax\} = x\dfrac{1}{F'(D)}\{\sin ax \text{ or } \cos ax\}$

Alternatively, $\sin ax$ and $\cos ax$ can be written in the form $\sin ax = \dfrac{e^{ixa} - e^{-ixa}}{2i}$ and $\cos ax = \dfrac{e^{aix} + e^{-aix}}{2}$, then find P.I. by the method of Type I.

**Example 17:** Solve $(D^4 + 2D^2 + 1)\,y = \cos x$.

**Solution:** The reduced equation is $(D^4 + 2D^2 + 1)\,y = 0$

Let $y = Ae^{mx}$ be a trial solution. Then the auxiliary equaiton is

$$m^4 + 2m^2 + 1 = 0 \text{ or } [(m^2 + 1)]^2 = 0 \text{ or } m = \pm i, \pm i$$

$\therefore$ C.F. $= (c_1 + c_2x)\cos x + (c_3 + c_4x)\sin x$ where $c_1$, $c_2$, $c_3$ and $c_4$ are arbitrary constants.

NOTES

$\therefore \qquad \text{P.I.} = \dfrac{1}{D^4 + 2D^2 + 1}\{\cos x\}$

$= x\dfrac{1}{4D^3 + 4D}\{\cos x\}$

$[\because \phi(D^2) = D^4 + 2D^2 + 1$

$\phi(-1^2) = 1 - 2 + 1 = 0$, then $\dfrac{1}{F(D)}\{f(x)\} = x\dfrac{1}{F'(D)}\{f(x)\}]$

$= \dfrac{x}{4}\dfrac{1}{D^3 + D}\{\cos x\} = \dfrac{x}{4}.\dfrac{x}{3D^2 + 1}\{\cos x\}$

$= \dfrac{x^2}{4}\;\dfrac{1}{3D^2+1}\{\cos x\} = \dfrac{x^2}{4}.\dfrac{\cos x}{-3+1} = -\dfrac{x^2}{8}\cos x$

Hence the general solution is

$y = \text{C.F.} + \text{P.I.}$

$= (c_1 + c_2x)\cos x + (c_3 + c_4x)\sin x - \dfrac{x^2}{8}\cos x.$

**Example 18:** Solve $(D^2 - 4)y = \sin 2x$.

**Solution:** The reduced equation is

$(D^2 - 4)y = 0$

Let $y = Ae^{mx}$ be a trial solution and then auxiliary equation is

$m^2 - 4 = 0 \Rightarrow m = \pm 2$

The complementary function is

$y = c_1 e^{2x} + c_2 e^{-2x}$ where $c_1, c_2$ are arbitrary constants.

The particular integral is

$y = \dfrac{1}{D^2 - 4}\{\sin 2x\} = \dfrac{1}{-2^2 - 4}\sin 2x$ [Replace $D^2$ by $-2^2$]

$= -\dfrac{1}{8}\sin 2x$

The general solution is $y = \text{C.F.} + \text{P.I.} = c_1e^{2x} + c_2e^{-2x} - \dfrac{1}{8}\sin 2x.$

**Example 19:** Solve $(3D^2 + 2D - 8)y = 5\cos x$.

**Solution:** The reduced equation is

$(3D^2 + 2D - 8)y = 0$

Let $y = Ae^{mx}$ be a trial solution and then the auxiliary equation is

$3m^2 + 2m - 8 = 0$ or $3m^2 + 6m - 4m - 8 = 0$

or $\qquad 3m(m + 2) - 4(m + 2) = 0$ or $(m + 2)(3m - 4) = 0$

NOTES

or $\quad m = -2, m = \frac{4}{3}$

$\therefore$ The complementary function is

$$y = c_1 e^{-2x} + c_2 e^{\frac{4}{3}x} \text{ when } c_1 \text{ and } c_2 \text{ are arbitrary constants.}$$

The particular integral is

$$y = \frac{1}{3D^2 + 2D - 8}\{5\cos x\} = 5\frac{1}{(3D-4)(D+2)}\{\cos x\}$$

$$= 5\frac{(3D+4)(D-2)}{(9D^2-16)(D^2-4)}\{\cos x\} = 5\frac{(3D+4)(D-2)}{[9(-1^2)-16][-1^2-4]}\{\cos x\}$$

[$D^2$ is replaced by $-1^2$ in the denominator] $\left[\frac{\psi(D)}{\phi(D^2)}\text{form}\right]$

$$= \frac{5}{(-25)(-5)}[3D^2 - 6D + 4D - 8]\{\cos x\} = \frac{1}{25}[3D^2 - 2D - 8]\cos x$$

$$= \frac{1}{25}\left(3\frac{d^2}{dx^2}(\cos x) - 2\frac{d}{dx}(\cos x) - 8\cos x\right)$$

$$= \frac{1}{25}[-3\cos + 2\sin x - 8\cos x] = \frac{1}{25}(2\sin x - 11\cos x)$$

The general solution is

$$y = \text{C.F.} + \text{P.I.}$$

$$= c_1 e^{-2x} + c_2 e^{4/3x} + \frac{1}{25}(2\sin x - 11\cos x).$$

**Type IV:** $F(D)\,y = x^n$, $n$ is a positive integer.

Here $\quad \text{P.I.} = \frac{1}{F(D)}\{x^n\} = [F(D)]^{-1}\{x^n\}$

In this case, $[F(D)]^{-1}$ is expanded in a binomial series in ascending powers of $D$ upto $D^n$ and then operate on $x^n$ with each term of the expansion. The terms in the expansion beyond $D^n$ need not be considered, since the result of their operation on $x^n$ will be zero.

**Example 20:** Solve $D^2(D^2 + D + 1)y = x^2$.

**Solution:** The reduced equation is

$$D^2(D^2 + D + 1)y = 0 \qquad (1)$$

Let $y = Ae^{mx}$ be a trial solution of Equation (2) and then the auxiliary equation is

$$m^2(m^2 + m + 1) = 0$$

$$\therefore\ m = 0, 0 \text{ and } m = \frac{-1 \pm \sqrt{1-4}}{2} = \frac{-1 \pm \sqrt{-3}}{2} = \frac{-I \pm \sqrt{3}i}{2}$$

$\therefore$ The complementary function is

$$y = (c_1 + c_2 x)\, e^{0 \cdot x} + e^{-\frac{1}{2}x}\left(c_3 \cos\frac{\sqrt{3}}{2}x + c_4 \sin\frac{\sqrt{3}}{2}x\right)$$

$$= c_1 + c_2 x + e^{-\frac{1}{2}x}\left(c_3 \cos\frac{\sqrt{3}}{2}x + c_4 \sin\frac{\sqrt{3}}{2}x\right)$$

where $c_1, c_2, c_3, c_4$ are the arbitrary constant.

The particular integral is

$$y = \frac{1}{D^2(D^2 + D + 1)}\{x^2\} = \frac{1}{D^2}(1 + D + D^2)^{-1}\{x^2\}$$

$$= \frac{1}{D^2}\{1 - (D + D^2) + (D + D^2)^2 - (D + D^2)^3 + ...\}\{x^2\}$$

$$= \frac{1}{D^2}\{1 - (D + D^2) + (D^2 + 2D^3 + D^4) - (D + D^2)^3 + ...\}\{x^2\}$$

$$= \frac{1}{D^2}\{x^2 - (2x + 2) + (2) + 0\}$$

$$= \frac{1}{D^2}\{x^2 - 2x\} = \frac{1}{D}\left\{\frac{x^3}{3} - x^2\right\} = \frac{x^4}{12} - \frac{x^3}{3}$$

The general solution is $y = \text{C.F.} + \text{P.I.}$

$$= c_1 + c_2 x + e^{-x/2}\left(c_3 \cos\frac{\sqrt{3}}{2}x + c_4 \sin\frac{\sqrt{3}}{2}x\right) + \frac{x^4}{12} - \frac{x^3}{3}.$$

**Example 21:** Solve $(D^2 + 4)y = x \sin^2 x$.

**Solution:** The reduced equation is

$$(D^2 + 4)\, y = 0$$

The trial solution $y = A\, e^{mx}$ gives the auxiliary equation as

$$m^2 + 4 = 0,\ m = \pm\, 2i$$

The complementary function is $y = c_1 \cos 2x + c_2 \sin 2x$

The particular integral is $y = \dfrac{1}{D^2 + 4}\{x \sin^2 x\}$

$$= \frac{1}{D^2 + 4}\left\{\frac{x}{2}(1 - \cos 2x)\right\} = \frac{1}{D^2 + 4}\left\{\frac{x}{2} - \frac{x}{2}\cos 2x\right\}$$

$$= \frac{1}{D^2 + 4}\left\{\frac{x}{2}\right\} - \frac{1}{D^2 + 4}\left\{\frac{x}{2}\,\frac{(e^{2ix} + e^{-2ix})}{2}\right\}$$

$$= \frac{1}{4}\left(1 + \frac{D^2}{4}\right)^{-1}\left\{\frac{x}{2}\right\} - \frac{1}{4}\,\frac{e^{2ix}}{(D + 2i)^2 + 4}\{x\} - \frac{e^{-2ix}}{4(D - 2i)^2 + 4}\{x\}$$

**NOTES**

$$= \frac{1}{4}\frac{x}{2} - \frac{e^{2ix}}{4}\frac{1}{D^2+4Di-4+4}\{x\} - \frac{e^{-2ix}}{4}\frac{1}{D^2-4Di-4+4}\{x\}$$

$$= \frac{x}{8} - \frac{e^{2ix}}{4}\frac{1}{4Di\left(1+\frac{D}{4i}\right)}\{x\} - \frac{e^{-2ix}}{4\,.(-4Di)\left(1-\frac{D}{4i}\right)}\{x\}$$

$$= \frac{x}{8} - \frac{e^{2ix}}{4}.\frac{1}{4Di}\left(1+\frac{D}{4i}\right)^{-1}\{x\} - \frac{e^{-2xi}}{4(-4Di)}\left(1-\frac{D}{4i}\right)^{-1}\{x\}$$

$$= \frac{x}{8} - \frac{e^{2ix}}{4}.\frac{1}{4Di}\left(1-\frac{D}{4i}+\frac{D^2}{-16}\ldots\right)\{x\} - \frac{e^{-2xi}}{4(-4Di)}\left(1+\frac{D}{4i}+\ldots\right)\{x\}$$

$$= \frac{x}{8} - \frac{e^{2ix}}{4}.\frac{1}{4Di}\left(x-\frac{1}{4i}\right) + \frac{e^{-2xi}}{4\,.\,4Di}\left(x+\frac{1}{4i}\right)$$

$$= \frac{x}{8} - \frac{e^{2ix}}{2\,.\,8i}\left(\frac{x^2}{2}-\frac{x}{4i}\right) + \frac{e^{-2xi}}{2\,.\,8i}\left(\frac{x^2}{2}+\frac{x}{4i}\right)$$

$$= \frac{x}{8} - \frac{x^2}{2\,.\,8}\left(\frac{e^{2ix}-e^{-2xi}}{2i}\right) + \frac{x}{2\,.16\,.\,i^2}\left(\frac{e^{2ix}+e^{-2xi}}{2}\right)$$

$$= \frac{x}{8} - \frac{x^2}{2.8}\sin 2x - \frac{x}{2.16}\cos 2x$$

$$= \frac{x}{8} - \frac{x^2}{16}\sin 2x - \frac{x}{32}\cos 2x$$

Hence the general solution is $y = \text{C.F.} + \text{P.I.}$

$$= c_1 \cos 2x + c_2 \sin 2x + \frac{x}{8} - \frac{x^2}{16}\sin 2x - \frac{x}{32}\cos 2x.$$

**Example 22:** Solve $(D^4 + D^3 - 3D^2 - 5D - 2)\,y = 3xe^{-x}$.

**Solution:** The reduced equation is

$$(D^4 + D^3 - 3D^2 - 5D - 2)\,y = 0 \qquad (1)$$

The trial solution $y = Ae^{mx}$ gives the auxiliary equation as

$m^4 + m^3 - 3m^2 - 5\,m - 2 = 0$

or $m^4 + m^3 - 3m^2 - 3\,m - 2m - 2 = 0$

or $m^3\,(m + 1) - 3m\,(m + 1) - 2\,(m + 1)$

or $(m + 1)\,(m^3 - 3m - 2) = 0$ or $(m + 1)\,\{m^3 + m^2 - m^2 - m - 2m - 2) = 0$

or $(m + 1)\,\{m^2\,(m + 1) - m\,(m + 1) - 2\,(m + 1)\} = 0$

or $(m + 1)\,(m + 1)\,(m^2 - m - 2) = 0$

or $(m + 1)^2\,(m^2 - 2m + m - 2) = 0$

or $(m + 1)^2\,(m + 1)\,(m - 2) = 0$

**NOTES**

$\therefore \quad m = -1, -1, -1, 2$

The complementary function is $y = (c_1 + c_2 x + c_3 x^2)\, e^{-x} + c_4 e^{2x}$.

The particular integral is

$$y = \frac{1}{(D+1)^3 (D-2)} \{3e^{-x} x\}$$

$$= 3e^{-x} \frac{1}{(D-1+1)^3 (D-3)} \{x\} = 3e^{-x} \frac{1}{D^3 (-3)(1-D/3)} \{x\}$$

$$= -e^{-x} \frac{1}{D^3} \left(1 - \frac{D}{3}\right)^{-1} \{x\} = -e^{-x} \frac{1}{D^3} \left(1 + \frac{D}{3} + \frac{D^2}{9} + \ldots\right) \{x\}$$

$$= -e^{-x} \frac{1}{D^3} \left(x + \frac{1}{3}\right) = -e^{-x} \frac{1}{D^2} \left(\frac{x^2}{2} + \frac{x}{3}\right) = -e^{-x} \frac{1}{D} \left(\frac{x^3}{6} + \frac{x^2}{6}\right)$$

$$= -e^{-x} \left(\frac{x^4}{24} + \frac{x^3}{18}\right)$$

The general solution is $y = \text{C.F.} + \text{P.I.}$

$$= (c_1 + c_2 x + c_3 x^2) + c_4 e^{2x} - e^{-x} \left(\frac{x^4}{24} + \frac{x^3}{18}\right).$$

**Type V:** ($a$) $F(D)\, y = xV$ where $V$ is a function of $x$.

Here $\quad \text{P.I.} = \dfrac{1}{F(D)} \{xV\} = \left\{x - \dfrac{1}{F(D)} F'(D)\right\} \dfrac{1}{F(D)} \{V\}.$

**Example 23:** Solve $(D^2 + 9)\, y = x \sin x$.

**Solution:** The reduced equation is $(D^2 + 9)\, y = 0$ (1)

The trial solution $y = Ae^{mx}$ gives the auxiliary equation as

$$m^2 + 9 = 0 \text{ or } m = \pm 3i$$

$\therefore$ C.F. $= c_1 \cos 3x + c_2 \sin 3x$ where $c_1$ and $c_2$ are arbitrary constants.

and P.I. $= \dfrac{1}{F(D)} \{x \sin x\}$ where $F(D) = D^2 + 9$

$$= \left\{x - \frac{1}{F(D)} F'(D)\right\} \frac{1}{F(D)} \{\sin x\}$$

$$= \left\{x - \frac{2D}{D^2 + 9}\right\} \frac{1}{D^2 + 9} \{\sin x\}$$

$$= \left\{x - \frac{2D}{D^2 + 9}\right\} \frac{\sin x}{-1 + 9} = \left\{x - \frac{2D}{D^2 + 9}\right\} \left\{\frac{\sin x}{8}\right\}$$

$$= \frac{x \sin x}{8} - \frac{1}{4} \frac{1}{-1 + 9} D\{\sin x\} = \frac{x \sin x}{8} - \frac{1}{32} \cos x$$

NOTES

Hence the general solution is

$$y = \text{C.F.} + \text{P.I.} = c_1 \cos 3x + c_2 \sin 3x + \frac{x \sin x}{8} - \frac{1}{32}\cos x$$

(*b*) $F(D)\, y = x^n V$ where $V$ is any function of $x$.

$$\text{Here P.I.} = \frac{1}{F(D)}\{f(x)\} = \frac{1}{F(D)}\{x^n V\} = \left\{x - \frac{F'(D)}{F(D)}\right\}^n \frac{1}{F(D)}\{V\}$$

**Example 24:** Solve $(D^2 - 1)y = x^2 \sin x$

**Solution:** The reduced equation is $(D^2 - 1)y = 0$ (2)

Let $y = Ae^{mx}$ be a trial solution. Then the auxiliary equation is

$$m^2 - 1 = 0 \text{ or } m = \pm 1$$

$\therefore$ C.F. $= c_1 e^x + c_2 e^{-x}$ where $c_1$ and $c_2$ are arbitrary constants.

$$\therefore \text{P.I.} = \frac{1}{F(D)}\{x^2 \sin x\} \text{ where } F(D) = D^2 - 1$$

$$= \left\{x - \frac{F'(D)}{F(D)}\right\}^2 \frac{1}{F(D)}\{\sin x\} = \left\{x - \frac{1}{D^2 - 1} 2D\right\}^2 \frac{1}{D^2 - 1}\{\sin x\}$$

$$= \left\{x - \frac{1}{D^2 - 1} 2D\right\}\left\{x - \frac{1}{D^2 - 1} 2D\right\}\left\{\frac{1}{-1^2 - 1}\sin x\right\}$$

$$= \left\{x - \frac{1}{D^2 - 1} 2D\right\}\left\{x - \frac{1}{D^2 - 1} 2D\right\}\{-1/2 \sin x\}$$

$$= \left\{x - \frac{1}{D^2 - 1} 2D\right\}\left\{-\frac{x}{2}\sin x + \frac{1}{D^2 - 1}\right\}\{\cos x\}$$

$$= \left\{x - \frac{1}{D^2 - 1} 2D\right\}\left\{-\frac{x}{2}\sin x - \frac{1}{2}\cos x\right\}$$

$$= -\frac{x^2}{2}\sin x - \frac{x}{2}\cos x + \frac{1}{D^2 - 1}\{D(x \sin x + \cos x)\}$$

$$= -\frac{x^2}{2}\sin x - \frac{x}{2}\cos x + \frac{1}{D^2 - 1}\{\sin x + x \cos x - \sin x\}$$

$$= -\frac{x^2}{2}\sin x - \frac{x}{2}\cos x + \frac{1}{D^2 - 1}\{x \cos x\}$$

$$\text{Again } \frac{1}{D^2 - 1}\{x \cos x\} = \left\{x - \frac{1}{D^2 - 1} 2D\right\}\frac{1}{D^2 - 1}\{\cos x\}$$

$$= \left\{x - \frac{1}{D^2 - 1} 2D\right\}\left\{\frac{1}{-1 - 1}\cos x\right\}$$

$$= -\frac{1}{2} x \cos x + \frac{1}{D^2 - 1}\{-\sin x\}$$

NOTES

$$= -\frac{1}{2}x\cos x - \frac{\sin x}{-1^2 - 1} = -\frac{1}{2}x\cos x + \frac{1}{2}\sin x$$

$$\therefore \text{P.I.} = -\frac{x^2}{2}\sin x - \frac{x}{2}\cos x - \frac{x}{2}\cos x + \frac{1}{2}\sin x$$

$$= -\frac{1}{2}x^2 \sin x - x\cos x + \frac{1}{2}\sin x$$

Hence the general solution is

$$y = \text{C.F.} + \text{P.I.} = c_1 e^x + c_2 e^{-x} - \frac{1}{2}x^2 \sin x - x\cos x + \frac{1}{2}\sin x.$$

### 9.5.1 Classification of Linear Partial Differential Equations of Second Order

Consider the following linear partial differential equation of the second order in two independent variables,

$$A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Fu = G$$

Where $A$, $B$, $C$, $D$, $E$, $F$ and $G$ are functions of $x$ and $y$.

This equation when converted to quasi-linear partial differential equation takes the form,

$$A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + f\left(x, y, u\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right) = 0$$

These equations are said to be of:

1. Elliptic type if $B^2 - 4AC < 0$
2. Parabolic type if $B^2 - 4AC = 0$
3. Hyperbolic type if $B^2 - 4AC > 0$

Let us consider some examples to understand this:

(i) $\frac{\partial^2 u}{\partial x^2} - 2x\frac{\partial^2 u}{\partial x \partial y} + x^2\frac{\partial^2 u}{\partial y^2} - 2\frac{\partial u}{\partial y} = 0$

$\Rightarrow u_{xx} - 2xu_{xy} + \text{x}2u_{yy} - 2u_y = 0$

Comparing it with the general equation we find that,

$A = 1$, $B = -2x$, $C = x^2$

Therefore

$B^2 - 4AC = (-2x)^2 - 4x^2 = 0$, $\forall$ $x$ and $y \neq 0$

So the equation is parabolic at all points.

(ii) $y^2u_{xx} + \text{x}^2u_{yy} = 0$

NOTES

Comparing it with the general equation we get,

$A = y^2, B = 0, C = x^2$

Therefore

$B^2 - 4AC = 0 - 4x^2y^2 < 0, \ \forall\ x$ and $y \neq 0$

So the equation is elliptic at all points.

(iii) $x^2u_{xx} - y^2u_{yy} = 0$

Comparing it with the general equation we find that,

$A = x^2, B = 0, C = -y^2$

Therefore

$B^2 - 4AC = 0 - 4x^2y^2 > 0, \ \forall\ x$ and $y \neq 0$

So the equation is hyperbolic at all points.

Following three are the most commonly used partial differential equations of the second order:

1. Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

This is equation is of elliptic type.

2. One-dimensional heat flow equation

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}$$

This equation is of parabolic type.

3. One-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

This is a hyperbolic equation.

## 9.6 HOMOGENEOUS AND NON-HOMOGENEOUS EQUATIONS WITH CONSTANT COEFFICIENTS

### Homogeneous Linear Equations with Constant Coefficients

Let $f(D, D')z = V(x, y)$ (9.39)

Then if

$$f(D, D') = A_0 D^n + A_1 D^{n-1} D' + A_2 D^{n-2} D'_2 + \cdots + A_n D'^n \qquad (9.40)$$

where $A_1, A_2, \cdots, A_n$ are constants.

Then Equation (9.39) is known as Homogeneous equation and takes the form

$$\left(A_0D^n + A_1D^{n-1}D' + A_2D^{n-2}D'^2 + \cdots + A_nD'^n\right)z = V(x, y) \quad (9.41)$$

**Complementary Function**

Consider the equation,

$$\left(A_0D^n + A_1D^{n-1}D' + A_2D^{n-2}D'^2 + \cdots + A_nD'^n\right)z = 0 \quad (9.42)$$

Let

$$z = \phi(y + mx) \quad (9.43)$$

be a solution of Equation (9.42)

Now $D^r z = m^r \phi^r (y + mx)$

$$D'^s z = \phi^{(e)}(y + mx)$$

and $D^r D'^s z = m^r \phi^{(r+s)}(y + mx)$

Therefore, on substituting Equation (9.43) in Equation (9.42), we get

$$\left(A_0m^n + A_1m^{n-s} + A_2m^{n-2} + \cdots + A_n\right)\phi^{(n)}(y + mx) = 0$$

which will be satisfied if

$$A_0m^n + A_1m^{n-1} + A_2m^{n-2} + \cdots + A_n = 0 \quad (9.44)$$

Equation (9.44) is known as the Auxiliary Equation.

Let $m_1, m_2, \cdots, m_n$ be the roots of the Equation (9.44),

Then the following three cases arise:

**Case I: Roots $m_1, m_2, \cdots, m_n$ are distinct.**

Part of C.F. corresponding to $m = m_1$ is

$$z = \phi_1(y + m_1x)$$

where '$\phi_1$' is an arbitrary function.

Part of C.F. corresponding to $m = m_2$ is

$$z = \phi_2(y + m_2x)$$

where $\phi_2$ is any arbitrary function.

Now since our equation is linear, so the sum of solutions is also a solution.

Therefore, our complimentary function becomes,

C.F. = $\phi_1(y + m_1x) + \phi_2(y + m_2x) + \ldots\ldots\ldots\ldots\ldots + \phi_n(y + m_nx)$

NOTES

**Case II: Roots are imaginary.**

Let the pair of complex roots of the Equation (9.44) be

$u \pm iv$

then the corresponding part of complimentary function is

$z = \phi_1(y + ux + ivx) + \phi_2(y + ux - ivx)$ …(9.45)

Let $y + ux = P$ and $vx = Q$

Then $z = \phi_1(P + iQ) + \phi_2(P - iQ)$

Or $z = (\phi_1 + \phi_2)P + (\phi_1 - \phi_2)iQ$

If $\phi_1 + \phi_2 = \xi_1$

And $\phi_1 - \phi_2 = \xi_2$

Then

$$\varphi_1 = \frac{1}{2}(\xi_1 + i\xi_2)$$

and

$$\varphi_2 = \frac{1}{2}(\xi_1 - i\xi_2)$$

Substituting these values in Equation (9.45), we get

$$z = \frac{1}{2}\xi_1(P + iQ) + \frac{1}{2}i\xi_2(P + iQ) + \frac{1}{2}\xi_1(P - iQ) - \frac{1}{2}i\xi_2(P - iQ)$$

or

$$z = \frac{1}{2}\{\xi_1(P + iQ) + \xi_1(P - iQ)\} + \frac{1}{2}i\{\xi_2(P + iQ) - \xi_2(P - iQ)\}$$

**Case III: Roots are repeated.**

Let $m$ be the repeated root of Equation (9.44).

Then we have,

$(D - mD')(D - mD')z = 0$

Putting $(D - mD')z = U$, we get (9.46)

$(D - mD')U = 0$ (9.47)

Since the equation is linear, it has the following subsidiary equations,

$$\frac{dx}{1} = \frac{dy}{-m} = \frac{dU}{0} \quad (9.48)$$

Two independent integrals of Equation (9.48) are

y + mx = constant

NOTES

and $\quad U = \text{constant}$

$\therefore \quad U = \phi(y + mx)$

is a solution of Equation (9.47) where $\phi$ is an arbitrary function.

Substituting in Equation (9.46)

$$\frac{\partial z}{\partial x} - m\frac{\partial z}{\partial y} = \phi(y + mx) \tag{9.49}$$

which has the following subsidiary equations,

$$\frac{dx}{1} = \frac{dy}{-m} = \frac{dz}{\phi(y + mx)}$$

Two independent integrals of Equation (9.46) are

$$y + mx = \text{constant}$$

and $\quad z = x\phi(y + mx) + \text{ constant}$

Therefore $\quad z = x\phi(y + mx) + \psi(y + mx) \qquad (9.50)$

is a solution of Equation (9.49) where $\psi$ is an arbitrary function.

Equation (9.50) is the part of C.F. corresponding to two times repeated root.

In general, if the root *m* is repeated *r* times, the corresponding part of C.F. is

$$z = x^{r-1}\phi_1(y + mx) + x^{r-2}\phi_2(y + mx) + \cdots + \phi_r(y + mx)$$

where $\phi_1, \phi_2, \cdots, \phi_r$ are arbitrary functions.

**Example 25:** Solve the equation, $(D^3 - 3D^2D' + 3DD'^2 - D'^3)z = 0$.

**Solution:** The A.E. of the given equation is

$$m^3 - 3m^2 + 3m - 1 = 0$$

or $\quad (m - 1)^3 = 0$

$\Rightarrow \quad m = 1, 1, 1$

$\therefore \quad \text{C.F.} = x^2\phi_1(y + x) + x\phi_2(y + x) + \phi_2(y + x)$.

**Non-Homogeneous Linear Equations with Constant Coefficients**

If all the terms on left hand side of Equation (9.39) are not of same degree then Equation (9.39) is said to be **Non-Homogeneous equation**. Equation is said to be **reducible** if the symbolic function $f(D, D')$ can be resolved into factors each of which is of first degree in *D* and *D'* and irreducible otherwise.

NOTES

For example, the equation

$$f(D,D')z=(D^2-D'^2+2D+1)z=(D+D'+1)(D-D'+1)z=x^2+xy$$

is reducible while the equation

$$f(D,D')z=(DD'+D'^3)z=D'(D+D'^2)z=\cos(x+2y)$$

is irreducible.

**Reducible Non Homogeneous Equations**

In the equation,

$$f(D,D')=(a_1D+b_1D'+c_1)(a_2D+b_2D'+c_2)\cdots(a_nD+b_nD'+c_n) \quad \ldots(9.51)$$

where *a*'s, *b*'s and *c*'s are constants.

The complementary function takes the form

$$(a_1D+b_1D'+c_1)(a_2D+b_2D'+c_2)\cdots(a_nD+b_nD'+c_n)z=0 \quad (9.52)$$

Any solution of the equation given by

$$(a_iD+b_iD'+c_i)z=0 \quad (9.53)$$

is a solution of the Equation (9.52)

Forming the Lagrange's subsidiary equations of Equation (9.53),

$$\frac{dx}{a_i}=\frac{dy}{b_i}=\frac{dz}{-c_iz} \quad (9.54)$$

The two independent integrals of Equation (9.54) are

$$b_ix-a_iy=\text{constant}$$

and $z=\text{constant}\ e^{-\frac{c_i}{a_i}x}$, if $a_i\neq 0$

or

$z=\text{constant}\ e^{-\frac{c_i}{b}y}$, if $b_i\neq 0$

Therefore

$$z=e^{-\frac{c_i}{a_i}x}\phi_i(b_ix-a_iy),\ \text{if } a_i\neq 0$$

or

$$z=e^{-\frac{c_i}{b_i}y}\psi_i(b_ix-a_iy) \quad \text{if } b_i\neq 0$$

is the general solution of Equation (9.53). Here $\varphi_i$ and $\psi_i$ are arbitrary functions.

NOTES

**Example 26:** Solve the differential equations

$$\left(D^2 - D'^2 - 3D + 3D'\right)z = 0.$$

**Solution:** The equation can also be written as

$$(D - D')(D + D' - 3)z = 0$$

$$\therefore \quad C.F. = \phi_1(y + x) + e^{3x}\phi_2(x - y)$$

Or

$$\psi_1(y + x) + e^{3y}\psi_2(x - y)$$

**When the Factors are Repeated**

Let the factor is repeated two times and is given by,

$$(aD + bD' + c)$$

Consider the equation

$$(aD + bD' + c)(aD + bD' + c)z = 0 \tag{9.55}$$

Put $(aD + bD' + c)z = U$ (9.56)

Then the Equation (9.51) reduces to

$$(aD + bD' + c)U = 0 \tag{9.57}$$

General solution of Equation (9.57) is

$$U = e^{-\frac{c}{a}x}\,\phi(bx - ay) \text{ if } a \neq 0 \tag{9.58}$$

Or

$$U = e^{-\frac{c}{b}y}\,\psi(bx - ay) \text{ if } b \neq 0 \tag{9.59}$$

Substituting Equation (9.58) in Equation (9.56), we obtain

$$(aD + bD' + c)z = e^{-\frac{c}{a}x}\,\phi(bx - ay) \tag{9.60}$$

The subsidiary equations are,

$$\frac{dx}{a} = \frac{dy}{b} = \frac{dz}{e^{-\frac{c}{a}x}\varphi(bx - ay) - cz} \tag{9.61}$$

The two independent integrals of Equation (9.61) are given by

$$bx - ay = \text{constant} = \lambda \tag{9.62}$$

and $$\frac{dz}{dx}+\frac{c}{a}z=\frac{1}{a}e^{-\frac{c}{a}x}\varphi(bx-ay)=\frac{1}{a}e^{-\frac{c}{a}x}\varphi(\lambda) \quad (9.63)$$

**NOTES**

The Equation (9.63) being an ordinary linear equation has the following solution:

$$ze^{\frac{c}{a}x}=\frac{1}{a}x\varphi(\lambda)+\text{constant}$$

or $$ze^{\frac{c}{a}x}=\frac{1}{a}x\varphi(bx-ay)+\text{constant}$$

Therefore, general solution of Equation (9.60) is

$$z=\frac{x}{a}e^{-\frac{c}{a}x}\phi(bx-ay)+\phi_1(bx-ay)e^{-\frac{c}{a}x}$$

$$=e^{-\frac{c}{a}x}\{x\phi_2(bx-ay)+\phi_1(bx-ay)\} \quad \ldots(9.64)$$

where $\phi_1$ and $\phi_2$ are arbitrary functions.

Similarly from Equations (9.59) and (9.56), we get

$$z=e^{-\frac{c}{b}y}\{y\psi_2(bx-ay)+\psi_1(bx-ay)\}$$

where $\psi_1$ and $\psi_2$ are arbitrary functions.

In general, for *r* times repeated factor, $(aD+bD'+c)$

$$z=e^{-\frac{c}{a}x}\sum_{i=1}^{r}x^{i-1}\phi_i(bx-ay) \qquad \text{if } a\neq 0$$

Or

$$z-e^{-\frac{c}{b}y}\sum_{i=1}^{r}y^{i-1}\psi_i(bx-ay) \qquad \text{if } b\neq 0$$

where $\varphi_1,\varphi_{2,},\ldots,\varphi_r$ and $\psi_1,\psi_2,\ldots,\psi_r$ are arbitrary functions.

**Example 27:** Solve the differential equation,

$$(2D-D'+4)(D+2D'+1^2z=0)$$

**Solution:** C.F. corresponding to the factor $(2D-D'+4)$ is

$$e^{4y}\phi(x+2y)$$

C.F. corresponding to the factor $(D=2D'+1)^2$ is

$$e^{-x}\{x\phi_2(2x-y)+\phi_1(2x-y)\}$$

Hence C.f. $= e^{4y}\phi(x+2y)+e^{-x}\{x\phi_2(2x-y)+\phi_1(2x-y)\}$

### Irreducible Non-Homogeneous Equations

NOTES

For solving the equation

$$f(D,D')z=0 \quad (9.65)$$

Substitute $z = ce^{ax+by}$ where $a$, $b$ and $c$ are constants (9.66)

Now $D^r z = ca^r e^{ax+by}$

$$D^r D'^a z = ca^r b^s e^{ax+by}$$

and $D'^s z = cb^s e^{ax+by}$

Substituting Equation (9.66) in Equation (9.65), we get,

$$cf(a,b)e^{ax+by} = 0$$

which will hold if

$$f(a, b) = 0 \quad (9.67)$$

For any selected value of $a$ (or $b$) Equation (9.67) gives one or more values of $b$ (or $a$). Thus there exists infinitely many pairs of numbers $(a_i, b_i)$ satisfying Equation (9.67).

Thus

$$z = \sum_{i=1}^{\infty} c_i e^{a_i x + b_i y} \quad (9.68)$$

where $f(a_i, b_i) = 0 \ \forall \ i$, is a solution of the Equation (9.65),

If

$$f(D,D') = (D + hD' + k)g(D,D') \quad (9.69)$$

then any pair $(a, b)$ such that

$$a + hb + k = 0 \quad (9.70)$$

satisfies Equation (9.67). There are infinite number of such solutions.

From Equation (9.70)

$$a = -(hb + k)$$

Thus

$$z = \sum_{i=1}^{\infty} c_i e^{-(hb_i + k)x + b_i y}$$

$$= e^{-kx}\sum_{i=1}^{\infty} c_i e^{b_i(y-hx)} \quad (9.71)$$

NOTES

is a part of C.F. corresponding to a linear factor $(D + hD' + k)$ given in Equation (9.69).

Equation (9.71) is equivalent to

$$e^{-kx}\phi(y - hx)$$

where '$\phi$' is an arbitrary function.

Equation (9.68) is the general solution if $f(D, D')$ has no linear factor otherwise general solution will be composed of both arbitrary functions and partly arbitrary constants.

**Example 28:** Solve the differential equation $(2D^4 + 3D^2D' + D'^2)z = 0$.

**Solution:** The given equation is equivalent to

$$(2D^2 + D')(D^2 + D')z = 0$$

C.F. corresponding to the first factor

$$= \sum_{i=1}^{\infty} c_i e^{a_i x + b_i y}$$

where $a_i$ and $b_i$ are related by

$$2a_i^2 + b_i = 0$$

or $$b_i = -2a_i^2$$

Therefore, part of C.F. corresponding to the first factor

$$\sum_{i=1}^{\infty} d_i e^{e_i(x - e_i y)}$$

where $e_i$ and $d_i$ are arbitrary constants.

$$\therefore \quad \text{C.F.} = \sum_{i=1}^{\infty} c_i e^{a_i(x-2a_i y)} + \sum_{i=1}^{\infty} d_i e^{e_i(x-e_i y)}$$

**Particular Integral**

In the equation,

$$f(D, D')z = V(x, y) \quad \ldots(9.72)$$

$f(D, D')$ is a non homogeneous function of $D$ and $D'$.

$$\text{P.I.} = \frac{1}{f(D, D')}V(x, y) \quad \ldots(9.73)$$

Here if $V(x, y)$ is of the form $e^{ax+by}$ where '*a*' and '*b*' are constants then we use the following theorem to evaluate the particular integral:

**Theorem 9.1:** If $f(a,b) \neq 0$, then

$$\frac{1}{f(D,D')}e^{ax+by} = \frac{1}{f(a,b)}e^{ax+by}$$

**Proof:** By differentiation

$$D^r D'^s e^{ax+by} = a^r b^s e^{ax+by}$$

$$D^r e^{ax+by} = a^r e^{ax+by}$$

$$D'^s e^{ax+by} = b^s e^{ax+by}$$

$$\therefore \quad f(D,D')e^{ax+by} = f(a,b)e^{ax+by}$$

$$e^{ax+by} = f(a,b)\frac{1}{f(D,D')}e^{ax+by}$$

Dividing the above equation by *f*(*a*, *b*)

$$\frac{1}{f(a,b)}e^{ax+by} = \frac{1}{f(D,D')}e^{ax+by}$$

or $$\frac{1}{f(D,D')}e^{ax+by} = \frac{1}{f(a,b)}e^{ax+by}$$

**Example 29:** Solve the equation $(D^2 - D'^2 - 3D + 3D')z = e^{x-2y}$

**Solution:** The given equation is equivalent to

$$(D - D')(D + D' - 3)z = e^{x-2y}$$

$$\text{C.F.} = \phi_1(y + x) + e^{3x}\phi_2(y - x)$$

$$\text{P.I.} = \frac{1}{(D - D')(D + D' - 3)}e^{x-2y}$$

$$= -\frac{1}{12}e^{x-2y}$$

Therefore, $z = \varphi_1(y + x) + e^{3x}\varphi_2(y - x) - \frac{1}{12}e^{x-2y}$

But in case *V*(*x*, *y*) is of the form $e^{ax+by}\phi(x, y)$ where '*a*' and '*b*' are constants then following theorem is used to evaluate the particular integral:

**NOTES**

**Theorem 9.2:** If $\phi(x,y)$ is any function, then

$$\frac{1}{f(D,D')}e^{ax+by}\phi(x,y)=e^{ax+by}\frac{1}{f(D+a,D'+b)}\phi(x,y)$$

**Proof:** From Leibnitz's theorem for successive differentiation, we have

$$D^r\{e^{ax+by}\phi(x,y)\}=e^{ax+by}\{D^r\phi(x,y)+{}^r c_1 a.D^{r-1}\phi(x,y)$$

$$+{}^r c_2 a^2 d^{r-2}\phi(x,y)+\cdots+{}^r c_r a^r\phi(x,y)$$

$$= e^{ax+by}\{D^r+{}^r c_1 D^{r-1}+{}^r c_2 a^2 D^{r-2}+\cdots+{}^r c^r a^r\}\phi(x,y)$$

$$= e^{ax+by}(D+a)^r\phi(x,y).$$

Similarly

$$D'^s\{e^{ax+by}\varphi(x,y)\}=e^{ax+by}(D'+b)^s\varphi(x,y)$$

and $D^r D'^s\{e^{ax+by}\varphi(x,y)\}=D^r[e^{ax+by}(D'+b)\varphi(x,y)]$

$$= e^{ax+by}(D+a)^r(D'+b)^s\phi(x,y)$$

So $f(D,D')\{e^{ax+by}\phi(x,y)\}=e^{ax+by}f(D+a,D'+b)\phi(x,y)$ (9.74)

Put $f(D+a,D'+b)\phi(x,y)=\psi(x,y)$

$\therefore$ $\phi(x,y)=\dfrac{1}{f(D+a,D'+b)}\psi(x,y)$

Substituting in Equation (9.74), we get

$$f(D,D')\left\{e^{ax+by}\frac{1}{f(D+a,D'+b)}\psi(x,y)\right\}=e^{ax+by}\psi(x,y)$$

Operating on the equation by $\dfrac{1}{f(D,D')}$

$$e^{ax+by}\frac{1}{f(D+a,D'+b)}\psi(x,y)=\frac{1}{f(D,D')}\{e^{ax+by}\psi(x,y)\}$$

Replacing $\psi(x,y)$ by $\phi(x,y)$, we have

$$\frac{1}{f(D,D')}(e^{ax+by}\phi(x,y))=e^{ax+by}\frac{1}{f(D+a,D'+b)}\phi(x,y)$$

NOTES

**Example 30:** Solve $\left(D^2 - D'^2 - 3D + 3D'\right)z = xv + e^{x+av}$.

**Solution:** The given equation is equivalent to,

$$(D - D')(D + D' - 3)y = xy + e^{x+2y}$$

$$\text{C.F.} = \phi_1(y + x) + e^{3x}\phi_2(x - y)$$

$$\text{P.I.} = \frac{1}{(D - D')(D + D' - 3)}xy + \frac{1}{(D - D')(D + D' - 3)}e^{x+2y}$$

$$= -\frac{1}{3D}\left\{1 - \frac{D'}{D}\right\}^{-1}\left\{1 - \frac{D + D'}{3}\right\}^{-1}xy$$

$$+ e^{x+2y}\frac{1}{(D + 1 - D' - 2)(D + 1 + D' + 2 - 3)}1.$$

$$= -\frac{1}{3D}\left\{1 + \frac{D'}{D} + \frac{D'^2}{D^2} + \cdots\right\}\left\{1 + \frac{D + D'}{3} + \frac{2}{9}DD' + \cdots\right\}xy + e^{x+2y}$$

$$\frac{1}{(D - D' - 1)(D + D')}.1$$

$$= -\frac{1}{3D}\left\{1 + \frac{D'}{D} + \frac{D'^2}{D^2} + \cdots\right\}\left\{xy + \frac{x + y}{3} + \frac{2}{9}.\right\} + e^{x+2y}\frac{1}{(-1)(D + D')}.1$$

$$= -\frac{1}{3D}\left\{xy + \frac{2}{3}x + \frac{x^2}{2} + \frac{1}{3}y + \frac{2}{9}\right\} - xe^{x+2y}$$

$$= -\frac{1}{3}\left\{\frac{x^2y}{2} + \frac{x^2}{3} + \frac{x^3}{6} + \frac{1}{3}xy + \frac{2}{9}x\right\} - xe^{x+2y}.$$

$$\therefore \quad z = \phi_1(y + x) + e^{3x}\phi_2(x - y) - xe^{x+2y} - \frac{1}{6}x^2y - \frac{1}{9}x^2 - \frac{x^3}{18} - \frac{1}{9}xy - \frac{2}{27}x$$

**Example 31:** Solve $\left(D^2 - DD' + D' - 1\right)z = \cos(x + 2y) + e^y + xy + 1$.

**Solution:** Equation is equivalent to

$$(D - 1)(D - D' + 1)z = \cos(x + 2y) + e^y + xy + 1$$

Complementary Function $= e^x\varphi_1(y) + e^y\varphi_2(x + y)$.

Particular integral corresponding to $\cos(x + 2y)$ is

**NOTES**

$$\frac{1}{D^2 - DD' + D' - 1}\cos(x+2y)$$

$$= \frac{1}{(-1)-(-2)+D'-1}\cos(x+2y)$$

$$= \frac{1}{D'}\cos(x+2y)$$

$$= \frac{1}{2}\sin(x+2y)$$

Corresponding to $e^y$, the particular integral is

$$= \frac{1}{D^2 - DD' + D' - 1}e^y$$

$$= \frac{1}{D'-1}e^y$$

$$= e^y.\frac{1}{D'}.1$$

$$= ye^y.$$

Particular Integral corresponding to the part $(xy+1)$ is

$$= \frac{1}{(D-1)(D-D'+1)}(xy+1)$$

$$-\{1-D\}^{-1}\{1+(D-D')\}^{-1}(xy+1)$$

$$= -\{1+D+D^2+.....\}\{1-(D-D')+(D-D')^2-.....\}(xy+1)$$

$$= -\{1+D+D^2+.....\}\{(xy+1)-(y-x)-2\}$$

$$= -\{1+D+D^2+.....\}(xy-y+x-1)$$

$$= -\{(xy-y+x-1)+(y+1)\}$$

$$= -(xy+x)$$

$$= -x(y+1)$$

$$\therefore \quad z = e^x\varphi_1(y) + e^y\varphi_2(x+y) + \frac{1}{2}\sin(x+2y) + ye^y - x(y+1)$$

## 9.7 PARTIAL DIFFERENTIAL EQUATIONS REDUCIBLE TO EQUATIONS WITH CONSTANT COEFFICIENTS

NOTES

The equation,

$$f(xD, yD')z = V(x, y)$$

where $f(xD, yD') = \sum_{r,s} c_{rs} x^r y^s D^r D'^s, c_{rs} = \text{constant.}$ (9.75)

is reduced to linear partial differential equation with constant coefficients by the following substitution:

$u = \log x, v = \log y$ (9.76)

By substitution of Equation (9.76)

$$xD = x\frac{\partial}{\partial x}$$

$$= x\frac{\partial}{\partial u}\frac{\partial u}{\partial x}$$

$$= \frac{\partial}{\partial u} = d(\text{say})$$

And

$$x^2D^2 = x^2D\left(\frac{1}{x}\frac{\partial}{\partial u}\right)$$

$$= x^2\left(-\frac{1}{x^2}\frac{\partial}{\partial u} + \frac{1}{x^2}\frac{\partial^2}{\partial u^2}\right)$$

$$= \frac{\partial^2}{\partial u^2} - \frac{\partial}{\partial u}$$

$$= d(d-1)$$

Therefore,

$$x^rD^r = d(d-1)(d-2).....(d-\overline{r-1})$$

and $y^sD'^s = d'(d'-1)(d'-2)...(d'-\overline{s-1})$

Hence $f(xD, yD') = \sum c_{rs} d(d-1).....(d-\overline{r-1})\, d'(d'-1).....(d'-\overline{s-1})$

$$= g(d, d')$$

**NOTES**

Here the coefficients in $g(d, d')$ are constants.

Thus by substitution Equation (9.75) is reduced to

$$g(d,d')z = V(e^u, e^v)$$

Or $g(d,d')z = U(u,v)$ (9.77)

Equation (9.77) can be solved by methods that have been described for solving partial differential equations with constant coefficients.

**Example 32:** Solve the differential equation,

$$(x^2D^2 - 4xyDD' + 4y^2D'^2 + 6yD')z = x^3y^4$$

**Solution:** Put $u = \log x$

$v = \log y$

The given equation can be reduced to

$$\{d(d-1) - 4dd' + 4d'(d'-1) + 6d'\}z = e^{3u+4v}$$

or $\{(d-2d')^2 - (d-2d')\}z = e^{3u+4v}$

or $(d-2d')(d-2d'-1)z = e^{3u+4v}$

The complementary function is $\phi_1(2u+v) + e^u\phi_2(2u+v)$

$= \phi_1(\log x^2y) + x\phi_2(\log x^2y)$

$= \psi_1(x^2y) + x\psi_2(x^2y)$

And the particular integral is $\dfrac{1}{(d-2d')(d-2d'-1)}e^{3u+2v}$

$= \dfrac{1}{30}e^{3u+4v}$

$= \dfrac{1}{30}x^3y^4$

$\therefore$ $z = \psi_1(x^2y) + x\psi_2(x^2y) + \dfrac{1}{30}x^3y^4$.

**Example 33:** Find the solution of, $(x^2D^2 - y^2D'^2 - yD' + xD)z = 0$

**Solution:** Put $u = \log x$

$v = \log y$

The given differential can be reduced to

NOTES

$$\{d(d-1)-d'(d'-1)-d'+d\}z=0$$

$$\Rightarrow \quad (d^2-d'^2)z=0$$

A.E. is

$$m^2-1=0$$

$$\Rightarrow \quad m=1,-1$$

$$\Rightarrow \quad z=\phi_1(v+u)+\phi_2(v-u)$$

$$=\phi_1(\log xy)+\phi_2\left(\log\frac{y}{x}\right)$$

$$=\Psi_1(xy)+\Psi_2\left(\frac{y}{x}\right).$$

**Example 34:** Determine the solution of the following equation: $(x^2D^2+2xyDD'+y^2D'^2)z+nz=n(xD+yD')z+x^2+y^2+x^3$

**Solution:** Put $u=\log x$

$$v=\log y$$

The Equation reduces to

$$\{d(d-1)+2dd'+d'(d'-1)\}z-n(d+d')z+nz=e^{2u}+e^{2v}+e^{3u}$$

or

$$\{(d+d')^2-(d+d')\}z-n(d+d')z+nz=e^{2u}+e^{2v}+e^{3u}$$

or

$$\{(d+d')(d+d'-1)-n(d+d')+n\}z=e^{2u}+e^{2v}+e^{3u}$$

or

$$\{(d+d')^2-(n+1)(d+d')+n\}z=e^{2u}+e^{2v}+e^{3u}$$

or

$$(d+d'-n)(d+d'-1)z=e^{2u}+e^{2v}+e^{3u}$$

C.F. $= e^{nu}\phi_1(u-v)+e^{u}\phi_2(u-v)$

$$=x^n\psi_1\left(\frac{x}{y}\right)+x\psi_2\left(\frac{x}{y}\right)$$

$$\text{P.I.} = \frac{1}{(d+d'-n)(d+d'-1)}\left\{e^{2u}+e^{2v}+e^{3u}\right\}$$

NOTES

$$= \frac{1}{2-n}e^{2u}+\frac{1}{2-n}e^{2v}+\frac{1}{(3-n)^2}e^{3u}$$

$$= -\frac{x^2+y^2}{n-2}-\frac{1}{2}.\frac{1}{n-3}x^3$$

$$\therefore\ z = x^n\psi_1\left(\frac{x}{y}\right)+x\psi_2\left(\frac{x}{y}\right)-\frac{x^2+y^2}{n-2}-\frac{1}{2}\frac{x^3}{n-3}$$

**Example 35:** Solve $\left(x^2D^2-xyDD'-2y^2D'^2+xD-2yD'\right)z=\log\frac{y}{x}-\frac{1}{2}$

**Solution:** Put $u=\log x$

$v=\log y$

Our equation reduces to

$$\{d(d-1)-dd'-2d'(d'-1)+d-2d'\}z = v-u-\frac{1}{2}$$

$$\left(d^2-dd'-2d'^2\right)z = v-u-\frac{1}{2}$$

or $$(d-2d')(d+d')z = v-u-\frac{1}{2}$$

$$\text{C.F.} = \phi_1(2u+v)+\phi_2(u-v)$$

$$= \psi_1\left(x^2y\right)+\psi_2\left(\frac{x}{y}\right)$$

P.I. $$= \frac{1}{(d-2d')(d+d')}\left(v-u-\frac{1}{2}\right)$$

$$= \frac{1}{d-2d'}.\frac{1}{d}\left\{1-\frac{d'}{d}\cdots\right\}\left(v-u-\frac{1}{2}\right)$$

$$= \frac{1}{d-2d'}.\frac{1}{d}\left\{v-u-\frac{1}{2}-u\right\}$$

**NOTES**

$$= \frac{1}{d-2d'}.\left(uv-u^2-\frac{1}{2}u\right)$$

$$= \frac{1}{d}\left\{1+\frac{2d'}{d}+\frac{4d'^2}{d^2}+\cdots\right\}\left(uv-u^2-\frac{1}{2}u\right)$$

$$= \frac{1}{d}\left\{uv-u^2-\frac{1}{2}u+u^2\right\}$$

$$= \frac{u^2v}{2}-\frac{u^2}{4}$$

$$= \frac{1}{2}(\log x)^2\log y-\frac{1}{4}(\log x)^2$$

$$\therefore \qquad z=\psi_1\left(x^2y\right)+\psi_2\left(\frac{x}{y}\right)+\frac{1}{2}(\log x)^2\log y-\frac{1}{4}(\log x)^2.$$

**Example 36:** Solve the differential equation,

$$\left(x^2D^2+2xyDD'+y^2D'^2\right)z=\left(x^2+y^2\right)^{\frac{n}{2}}$$

**Solution:** Put $u=\log x$

$v=\log y$

The equation is reduced to $\{d(d-1)+2dd'+d'(d'-1)\}z=\left(e^{2u}+e^{2v}\right)^{\frac{n}{2}}$

or $\quad \left\{(d+d')^2-(d+d')\right\}z=\left(e^{2u}+e^{2v}\right)^{\frac{n}{2}}$

or $\quad (d+d')(d+d'-1)z=\left(e^{2u}+e^{2v}\right)^{\frac{n}{2}}$

C.F. $\quad =\phi_1(u-v)+e^u\phi_2(u-v)$

$$=\phi_1\left(\log\frac{x}{y}\right)+x\phi_2\left(\log\frac{x}{y}\right)$$

$$=\Psi_1\left(\frac{x}{y}\right)+x\Psi_2\left(\frac{x}{y}\right)$$

Particular Integral is $=\dfrac{1}{(d+d')(d+d'-1)}\left(e^{2u}+e^{2v}\right)^{\frac{n}{2}}$

NOTES

Substituting $Z = \frac{1}{d + d' - 1}\left(e^{2u} + e^{2v}\right)^{\frac{n}{2}}$

or $$\frac{\partial Z}{\partial u} + \frac{\partial Z}{\partial v} = Z + \left(e^{2u} + e^{2v}\right)^{\frac{n}{2}}$$

The subsidiary equations are $$\frac{du}{1} = \frac{dv}{1} = \frac{dZ}{Z + \left(e^{2u} + e^{2v}\right)^{\frac{n}{2}}}$$

Two independent integrals of Equation are given by

$$u - v = \text{constant} = a \text{ (say)}$$

and $$\frac{dZ}{dv} - Z = \left(e^{2u} + e^{2v}\right)^{\frac{n}{2}}$$

$$= e^{nv}(e^{2a} + 1)^{\frac{n}{2}}$$

Since this equation is linear, therefore

$$Ze^{-v} = \frac{e^{(n-1)v}}{(n-1)}\left(e^{2a} + 1\right)^{\frac{n}{2}}$$

$$\therefore \quad Z = \frac{e^{nv}}{n-1}\left(e^{2a} + 1\right)^{\frac{n}{2}}$$

$$= \frac{\left(e^{2u} + e^{2v}\right)^{\frac{n}{2}}}{(n-1)}$$

$$\therefore \quad \text{P.I.} = \frac{1}{d + d'}\left\{\frac{\left(e^{2u} + e^{2v}\right)^{\frac{n}{2}}}{n-1}\right\}$$

$$= \frac{1}{(n-1)}\int_{a=v-u}\left\{e^{2u} + e^{2a+2u}\right\}^{\frac{n}{2}} du$$

$$= \frac{1}{n-1}\left\{\int (e^{2a} + 1)^{\frac{n}{2}} \int e^{nu}\, du\right\}_{a=v-u}$$

$$= \frac{1}{n(n-1)}\left\{e^{nu}\left(e^{2a} + 1\right)^{\frac{n}{2}}\right\}_{a=v-u}$$

$$= \frac{1}{n(n-1)}\left(e^{2u} + e^{2v}\right)^{\frac{n}{2}}$$

$$= \frac{1}{n(n-1)}\left(x^2 + y^2\right)^{\frac{n}{2}}$$

NOTES

$$\therefore \quad z = \psi_1\left(\frac{x}{y}\right) + x\psi_2\left(\frac{x}{y}\right) + \frac{1}{n(n-1)}\left(x^2 + y^2\right)^{\frac{n}{2}}.$$

**Example 37:** Solve $\left(x^2D^2 - 2xyDD' + y^2D'^2 - xD + 3yD'\right)z = \frac{8y}{x}$

**Solution:** Put $\quad u = \log x$

$v = \log x$

Our Equation reduces to

$$\{d(d-1) - 2dd' + d'(d'-1) - d + 3d'\}z = 8e^{v-u}$$

or $\quad \{(d-d')^2 - 2(d-d')\}z = 8e^{v-u}$

or $\quad (d-d')(d-d'-2)z = 8e^{v-u}$

C.F. $\quad = \phi_1(u+v) + e^{2u}\phi_2(u+v)$

$= \psi_1(xy) + x^2\psi_2(xy)$

P.I. $\quad = 8.\frac{1}{(d-d')(d-d'-2)}e^{v-u}$

$= e^{v-u}$

$= \frac{y}{x}$

$$\therefore \quad z = \psi(xy) + x^2\psi_2(xy) + \frac{y}{x}.$$

**Example 38:** Solve $\left(x^2D^2 + 2xyDD' + y^2D'^2\right)z = x^m y^n$

**Solution:** Put $\quad u = \log x$

$v = \log y$

The equation reduces to

NOTES

$$\{d(d-1)+2dd'+d'(d'-1)\}z = e^{mu+nv}$$

or $$\{(d+d')^2-(d+d')\}z = e^{mu+nv}$$

or $$(d+d')(d+d'-1)z = e^{mu+nv}$$

$$\text{C.F.} = \phi_1(u-v)+e^u\phi_2(u-v)$$

$$= \psi_1\left(\frac{x}{y}\right)+x\psi_2\left(\frac{x}{y}\right)$$

$$\text{P.I.} = \frac{1}{(d+d')(d+d'-1)}e^{mu+nv}$$

$$= \frac{1}{(m+n)(m+n-1)}e^{mu+nv}$$

$$= \frac{1}{(m+n)(m+n-1)}x^m y^n$$

$$\therefore \quad z = \psi_1\left(\frac{x}{y}\right)+x\psi_2\left(\frac{x}{y}\right)+\frac{1}{(m+n)(m+n-1)}x^m y^n.$$

**Check Your Progress**

5. Write the general linear differential equation with constant coefficients.
6. What are the three types of second order partial differential equations?
7. What is the complementary function of the equation $(A_0D^n + A_1D^{n-1}D' + A_2D^{n-2}D'^2 + \cdots + A_nD'^n)z = 0$ if the roots are distinct?
8. When is a non-homogeneous equation said to be reducible?
9. Which mathematical function is used to reduce partial differential equations to equations with constant coefficients?

## 9.8 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The partial differential equation $Pp + Qq = R$, where $P, Q, R$ are functions of $x, y, z$ is called Lagrange's linear differential equation.
2. We have to make following assumptions:
   (a) The mass of the string for each unit length is constant ('homogeneous string'). The string is perfectly elastic and does not offer any resistance to bending.

NOTES

(b) The tension caused by stretching the string before fixing it at the ends is so large that the action of the gravitational force on the string can be neglected.

(c) The string performs small transverse motions in a vertical plane; that is, every particle of the string moves strictly vertically and so that the deflection and the slope at every point of the string always remain small in absolute value.

3. Each $u_n(x,t) = (B_n \cos\lambda_n t + B_n{}^* \sin\lambda_n t)\sin\frac{n\pi}{L}x$ represents a harmonic motion having the frequency $\lambda_n / 2\pi = cn/2L$ cycles per unit time. This motion is called the $n$th normal mode of the string.
4. Charpit's method is used to find the solution of most general partial differential equation of order one.
5. The linear differential equation with constant coefficients are of the form,

$$\frac{d^n y}{dx^n} + P_1\frac{d^{n-1}y}{dx^{n-1}} + P_2\frac{d^{n-2}y}{dx^{n-2}} + ..... + P_{n-1}\frac{dy}{dx} + P_n y = Q$$

Where $P_1, P_2, \ldots, P_n$ are constants and $Q$ is a function of $x$.
6. The three types of equations are the elliptic type, the parabolic type and the hyperbolic type.
7. Let $m_1, m_2, ..., m_n$ be the roots of the equation then C.F. = $\phi_1(y + m_1x) + \phi_2(y + m_2x) + \ldots\ldots\ldots\ldots + \phi_n(y + m_nx)$ where $\phi_i$'s are arbitrary functions.
8. The equation $f(D, D')z = V(x, y)$ is said to be reducible if the symbolic function $f(D, D')$ can be resolved into factors each of which is of first degree in $D$ and $D'$.
9. Logarithm function is used to reduce partial differential equations to equations with constant coefficients

## 9.9 SUMMARY

- Lagrange's equation can be solved by forming auxiliary equations and then finding two independent solutions of the auxiliary equations.
- Wave equation governs the motion of a violin string.
- Charpit's method is used to find the solution of most general partial differential equation of order one.
- The general solution of $F(D)y = Q$ consists of two parts:
  - o The complementary function which is the complete primitive of the reduced equation and is of the form

$y = c_1y_1 + c_2y_2 + \ldots + c_ny_n$ containing $n$ arbitrary constants.

o The particular integral which is a solution of $F(D)y = Q$ containing no arbitrary constant.

NOTES

- Second order partial differential equations can be classified as elliptic, parabolic or hyperbolic type.
- If $f(D, D2)z = V(x, y)$ and

  $f(D, D') = A_0D^n + A_1D^{n-1}D' + A_2D^{n-2}D_2' + \cdots + A_nD'^n$ where

  $A_1, A_2, \cdots, A_n$ are constants then the equation is known as homogeneous equation.
- The roots of homogeneous equation can be distinct, repeated or imaginary.
- If all the terms on left hand side of Equation $f(D, D')z = V(x, y)$ are not of same degree then equation is said to be non homogeneous equation. Equation is said to be reducible if the function $f(D, D')$ can be resolved into factors each of which is of first degree in $D$ and $D'$ and irreducible otherwise.
- The equation, $f(xD, yD')z = V(x, y)$ where

  $f(xD, yD') = \sum_{r,s} c_{rs}x^ry^sD^rD'^s, c_{rs}$ = constant is reduced to linear partial differential equation with constant coefficients by the substitution, $u = \log x$ and $v = \log y$

## 9.10 KEY WORDS

- **Partial differential equation:** Any equation which contains one or more partial derivatives is called a partial differential equation.
- **Fundamental mode**: The first normal mode is referred as the fundamental mode.

## 9.11 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. Define partial differential equations with suitable examples.
2. How will you identify the order of a partial differential equation?
3. Which equations are termed as singular integral?
4. How will you determine the degree of the partial differential equation?
5. What is a spectrum?

6. Define Wronskian of functions.
7. Give examples of parabolic, elliptic and hyperbolic type equations.
8. What is the difference between homogeneous and non homogeneous differential equations?

### Long-Answer Questions

1. Solve the following differential equations:

   a. $(3z-4y)p+(4x-2z)q=2y-3x$

   b. $x(z^2-y^2)p+y(x^2-z^2)q=z(y^2-x^2)$

2. How does the frequency of the fundamental mode of the vibrating string depend on the (a) Length of the string (b) On the mass per unit length (c) On the tension? What happens to that frequency if we double the tension?
3. Find $u(x, t)$ of the string of length $L = \pi$ when $c^2 = 1$, the initial velocity is zero, and the initial deflection is

   a. $0.01 \sin 3x$.

   b. $k\left(\sin x-\frac{1}{2}\sin 2x\right)$.

   c. $0.1x(\pi-x)$.

   d. $0.1x(\pi^2-x^2)$.

4. Find the deflection $u(x,t)$ of the string of length $L = \pi$ and $c^2=1$ for zero initial displacement and 'triangular' initial velocity $u_t(x,0)=(0.01x)$ if $0\leqq x\leqq\frac{1}{2}\pi$, $u_t(x,0)=0.01(\pi-x)$ if $\frac{1}{2}\pi\leqq x\leqq\pi$. (Initial conditions with $u_t(x,0)\neq 0$ are hard to realize experimentally).
5. Find solutions $u(x, y)$ of the following equations by separating variables.

   a. $u_x+u_y=0$.

   b. $u_x-u_y=0$.

   c. $y^2u_x-x^2u_y=0$.

   d. $u_x+u_y=(x+y)u$.

   e. $u_{xx}+u_{yy}=0$.

NOTES

f. $u_{xy} - u = 0$.

g. $u_{xx} - u_{yy} = 0$.

h. $xu_{xy} + 2yu = 0$.

6. Show that

a. The substitution of $u(x,t) = \sum_{n=1}^{\infty} G_n(t)\sin\frac{n\pi x}{L}$ ($L$ =length of the string) into the wave equation governing free vibrations leads to

$$\ddot{G}_n + \lambda_n^2 G = 0,\ \lambda_n = \frac{cn\pi}{L}.$$

b. Forced vibrations of the string under an external force $P(x, t)$ per unit length acting normal to the string are governed by the equation

$$u_{tt} = c^2 u_{xx} + \frac{P}{\rho}.$$

7. Find Complete Integrals of the following equations:

a. $p^2 + px + q = z$.

b. $p^2x + q^2y = z$.

c. $px + qy = z\sqrt{1+pq}$.

d. $p(1 + q^2) = q(z - a)$.

e. $pq + x(2y+1)p + (y^2 + y)q - (2y+1)z = 0$.

f. $(pq)(px + qy) = 1$.

g. $pxy + pq + qy = yz$.

h. $(p^2 + q^2)x = pz$.

i. $2(y + zq) = q(xp + yq)$.

8. Solve the equations:

a. $(\mathrm{D}^2 + \mathrm{DD}'^{s} - 1\mathrm{D}'^3)z = 0$.

b. $(\mathrm{D}^3 + 3\mathrm{D}^2\mathrm{D}' - 4\mathrm{D}'^3)z = 0$.

9. Solve the equations:

a. $(\mathrm{D}^2 + 2\mathrm{DD}' + \mathrm{D}'^2)z = 12xy$.

b. $(\mathrm{D}^2 - 2\mathrm{DD}' - 15\mathrm{D}'^2)z = 12xy$.

**NOTES**

c. $\left(D^2 - 6DD' - 9D'^2\right)z = 12x^2 + 16xy$.

d. $\left(D^3 - 7DD'^2 - 6D'^3\right)z = x^2 + xy^2 + y^3$.

e. $\left(D^2D' - 2DD'^2 + D'^3\right)z = \frac{1}{x^2}$.

10. Solve the equations:

a. $\left(D^2 - DD' - 2D'^2\right)z = x - y$.

b. $\left(D^2 - 3DD' + 2D'^2\right)z = x + y$.

c. $\left(4D^2 - 4DD' + D'^2\right)z = 16\log(x + 2y)$.

d. $\left(D^3 - 7DD'^2 - 6D'^3\right)z = \cos(x - y) + x^2 + xy^2 + y^3$.

e. $\left(D^3 - 7DD'^2 - 6D'^3\right)z = \sin(x + 2y) + e^{3x+y}$.

f. $\left(D^3 - 3DD'^2 + 2D'^3\right)z = \sqrt{x - 2y}$.

g. $\left(D^3 - 4D^2D' + 5DD'^2 - 2D'^3\right)z = e^{y+2x} + \sqrt{y + x}$.

11. Solve the equations:

a. $\left(D^3 - 3DD'^2 - 2D'^3\right)z = \cos(x + 2y)$.

b. $\left(D^2 + 5DD' + 5D'^2\right)z = x\sin(3x - 2y)$.

12. Solve the equations:

a. $\left(D^2 - Dd' - 2D'^2\right)z = (y - 1)e^x$.

b. $\left(D^3 - 3DD'^2 - 2D'^3\right)z = \cos(x + 2y) - e^y(3 + 2x)$.

13. Solve the equations:

a. $\left(DD' + D'^2 - 3D'\right)z = 0$.

b. $(2D + D' - 1)^2(D - 2D' + 2)^3 z = 0$.

14. Solve the equations:

a. $\left(2D^2 - D'^2 + D\right)z = 0$.

b. $\left(D^2 + DD' + D + D' + 1\right)z = 0$.

15. Solve the equations:

a. $(D - D' - 1)(D + D' - 2)z = e^{2x-y}$.

b. $\left(D^2 - D'\right)z = e^{x+y}$.

**NOTES**

16. Solve the equations:

    a. $\left(D^2 - DD' - 2D\right)z = \cos(3x + 4y)$.

    b. $\left(D^2 - D'\right)z = A\cos(lx + my)$, where *A*, *l*, *m* are constants.

17. Solve the equations:

    a. $(D = D' - 1)(D + 2D' - 3)z = 4 + 3x + 6y$.

    b. $\left(D^3 - DD'^2 - D^2 + DD'\right)z = \dfrac{x+2}{x^3}$.

    c. $\left(D^2 - D'\right)y = 2y - x^2$.

18. Solve the equations:

    a. $\left(D - D'^2\right)z = \cos(x - 3y)$.

    b. $(D + D' - 1)(D + D' - 3)(D + D')z = e^{x+y}\sin(2x + y)$.

    c. $\left(D^2 + DD' + D' - 1\right)z = 4\sin h\ x$.

    d. $\left(D^2D' + D'^2 - 2\right)z =^{2y} \sin 3x - e^{\infty}\cos 2y$.

19. Solve the equations:

    a. $\left(x^2D^3 - y^3D'^2\right)z = xy$.

    b. $\left(x^2D^2 + 2xyDD' + y^2D'^2\right)z = x^2y^2$.

    c. $\left(x^2D^2 - 2xyDD' - 3y^2D'^2 + xD - 3yD'\right)z = x^2y\cos\left(\log x^3\right)$.

20. Solve $\left(D^3 - 2D^2D' - DD'^2 + 2D'^3\right)z = e^{x+y}$.

21. Solve $\left(D^3 + D'^3 + D''^3 - 3DD'D''\right)u = x^3 = 3xyz$.

22. Solve the following equations:

    a. $r = x^2 e^y$.

    b. $x\,ys = 1$.

23. Solve the following equations:

    a. $t - xq = -\sin y - x\cos y$.

    b. $t - xq = x^2$.

    c. $yt - q = xy$.

24. Solve the following equations:

    a. $xr + ys + p = 10xy^3$.

b. $2yt - xs + 2q = 4yx^3$.

c. $z + r = x\cos(x + y)$.

25. Solve the differential equation, $r - 2yp + y^2 z = (y - 2)e^{2x+3y}$.

NOTES

## 9.12 FURTHER READINGS

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

# UNIT 10 ORDINARY DIFFERENTIAL EQUATIONS

NOTES

**Structure**



## 10.0 INTRODUCTION

In mathematics, an ordinary differential equation is a relation that contains functions of only one independent variable and one or more of their derivatives with respect to that variable. Ordinary differential equations are distinguished from partial differential equations, which involve partial derivatives of functions of several variables. Ordinary differential equations arise in many different contexts including geometry, mechanics, astronomy and population modelling. The Picard—Lindelöf theorem, Picard's existence theorem or Cauchy–Lipschitz theorem is an important theorem on existence and uniqueness of solutions to first-order equations with given initial conditions. The Picard method is a way of approximating solutions of ordinary differential equations. Originally, it was a way of proving the existence of solutions.

In this unit, you will study about the ordinary differential equations and local truncation error.

## 10.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the ordinary differential equations
- Analyse the local truncation error

## 10.2 ORDINARY DIFFERENTIAL EQUATIONS

Even though there are many methods to find an analytical solution of ordinary differential equations, for many differential equations solutions in closed form cannot be obtained. There are many methods available for finding a numerical solution for

differential equations. We consider the solution of an initial value problem associated with a first order differential equation given by,

NOTES

$$\frac{dy}{dx} = f(x, y) \tag{10.1}$$

With $$y(x_0) = y_0 \tag{10.2}$$

In general, the solution of the differential equation may not always exist. For the existence of a unique solution of the differential Equation (10.1), the following conditions, known as Lipshitz conditions must be satisfied,

(*i*) The function $f(x, y)$ is defined and continuous in the strip

$R: \; x_0 \le x \le b, \quad -\infty < y < \infty$

(*ii*) There exists a constant $L$ such that for any $x$ in $(x_0, b)$ and any two numbers $y$ and $y_1$

$$|f(x, y) - f(x, y_1)| \le L|y - y_1| \tag{10.3}$$

The numerical solution of initial value problems consists of finding the approximate numerical solution of $y$ at successive steps $x_1, x_2, ..., x_n$ of $x$. A number of good methods are available for computing the numerical solution of differential equations.

### Picard's Method of Successive Approximations

Consider the solution of the initial value problem,

$$\frac{dy}{dx} = f(x, y) \text{ with } y(x_0) = y_0$$

Taking $y = y(x)$ as a function of $x$, we can integrate the differential equation with respect to $x$ from $x = x_0$ to $x$, in the form

$$y = y_0 + \int_{x_0}^{x} f(x, y(x))\, dx \tag{10.4}$$

The integral contains the unknown function $y(x)$ and it is not possible to integrate it directly. In Picard's method, the first approximate solution $y^{(1)}(x)$ is obtained by replacing $y(x)$ by $y_0$.

Thus, $$y^{(1)}(x) = y_0 + \int_{x_0}^{x} f(x, y_0) dx \tag{10.5}$$

The second approximate solution is derived on replacing $y$ by $y^{(1)}(x)$. Thus,

$$y^{(2)}(x) = y_0 + \int_{x_0}^{x} f(x, y^{(1)}(x))\, dx \tag{10.6}$$

NOTES

The process can be continued, so that we have the general approximate solution given by,

$$y^{(n)}(x) = y_0 + \int_{x_0}^{x} f(x, y^{(n-1)}(x))dx, \text{ for } n = 2, 3... \quad (10.7)$$

This iteration formula is known as Picard's iteration for finding solution of a first order differential equation, when an initial condition is given. The iterations are continued until two successive approximate solutions $y^{(k)}$ and $y^{(k+1)}$ give approximately the same result for the desired values of $x$ up to a desired accuracy.

***Note:*** Due to practical difficulties in evaluating the necessary integration, this method cannot be always used. However, if $f(x, y)$ is a polynomial in $x$ and $y$, the successive approximate solutions will be obtained as a power series of $x$.

**Example 1:** Find four successive approximate solutions for the following initial value problem: $y' = x + y,$ with $y(0) = 1$, by Picard's method. Hence compute $y(0.1)$ and $y(0.2)$ correct to five significant digits.

**Solution:** We have, $y' = x + y,$ with $y(0) = 1.$

The first approximation by Picard's method is,

$$y^{(1)}(x) = y(0) + \int_0^x [x + y(0)]\, dx$$

$$\therefore \quad y^{(1)}(x) = 1 + \int_0^x (x+1) \quad dx = 1 + x + \frac{x^2}{2}$$

The second approximation is,

$$y^{(2)}(x) = 1 + \int_0^x (x + 1 + x + \frac{x^2}{2})dx = 1 + x + x^2 + \frac{x^3}{6}$$

Similarly, the third approximation is,

$$y^{(3)}(x) = 1 + \int_0^x (1 + 2x + x^2 + \frac{x^3}{6})dx$$

$$\therefore \quad y^{(3)}(x) = 1 + x + x^2 + \frac{x^3}{3} + \frac{x^4}{24}$$

The fourth approximation is,

$$y^{(4)}x = 1 + \int_0^x (1 + 2x + x^2 + \frac{x^3}{3} + \frac{x^4}{24})dx$$

$$\therefore \quad y^{(4)}(x) = 1 + x + x^2 + \frac{x^3}{3} + \frac{x^4}{12} + \frac{x^5}{120}$$

It is clear that successive approximations are easily determined as power series of $x$ having one degree more than the previous one. The value of $y(0.1)$ is given by,

$y\ (0.1) = 1 + 0.1 + (0.1)^2 + \frac{(0.1)^3}{3} + \frac{(0.1)^4}{4} + ... \approx 1.1103,$ correct to five significant digits.

Similarly, $y(0.2) = 1 + 0.2 + (0.2)^2 + \frac{(0.2)^3}{3} + \frac{(0.2)^4}{4} + \frac{(0.2)^5}{120} \approx 1.2431.$

**Example 2:** Find the successive approximate solution of the initial value problem, $y' = xy + 1,$ with $y\ (0) = 1$, by Picard's method.

**Solution:** The first approximate solution is given by,

$$y^{(1)}(x) = 1 + \int_0^x (x+1)dx = 1 + x + \frac{x^2}{2}$$

The second and third approximate solutions are,

$$y^{(2)}(x) = 1 + \int_0^x [x(1 + x + \frac{x^2}{2}) + 1]dx = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{8}$$

$$y^{(3)}(x) = 1 + \int_0^x [x(1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4}) + 1]dx = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{8} + \frac{x^5}{15} + \frac{x^6}{48}$$

**Example 3:** Compute $y\ (0.25)$ and $y\ (0.5)$ correct to three decimal places by solving the following initial value problem by Picard's method:

$$\frac{dy}{dx} = \frac{x^2}{1 + y^2}, y(0) = 0$$

**Solution:** We have $\frac{dy}{dx} = \frac{x^2}{1 + y^2}, y(0) = 0$

By Picard's method, the first approximation is,

$$y^{(1)}(x) = 0 + \int_0^x \frac{x^2}{1+0} dx = \frac{x^3}{3}$$

The second approximate solution is,

$$y^{(2)}(x) = \int_0^x \frac{x^2}{1 + [y^{(1)}(x)]^2} dx$$

$$= \int_0^x \frac{x^2}{1 + \frac{x^6}{9}} dx = \tan^{-1} \frac{x^3}{3}$$

For $x = 0.25, \quad y^{(1)} \quad (0.25) = \frac{(0.25)^2}{3} = 0.0052$

$$y^{(2)}\quad(0.25) = \tan^{-1}\ \frac{(0.25)^2}{3} \approx 0.0052$$

$$\therefore y\quad(0.25) = 0.005, \text{Correct to three decimal place.}$$

**NOTES**

Again, for $x = 0.5$, $$y^{(1)}(0.5) = \frac{(0.5)^2}{3} = 0.083333$$

$$y^{(2)}(0.5) = \tan^{-1}\frac{(0.5)^3}{3} = 0.0416$$

Thus, correct to three decimal places, $y(0.5) = 0.042$.

***Note:*** For this problem we observe that, the integral for getting the third and higher approximate solution is either difficult or impossible to evalute, since

$$y^{(3)}(x) = \int_0^x \frac{x^2}{1+\left(\tan^{-1}\frac{x^3}{3}\right)^2} \text{ is not integrable.}$$

**Example 4:** Use Picard's method to find two successive approximate solutions of the initial value problem,

$$\frac{dy}{dx} = \frac{y-x}{y+x},\ y(0) = 1$$

**Solution:** The first approximate solution by Picard's method is given by,

$$y^{(1)}(x) = y_0 + \int_0^x f(x, y_0)dx$$

$$\therefore \qquad y^{(1)}(x) = 1 + \int_0^x \frac{1-x}{1+x}dx = 1 + \int_0^x \frac{2-(1+x)}{1+x}dx$$

$$\therefore \qquad y^{(1)}(x) = 1 + 2\log_e|1+x| - x$$

The second approximate solution is given by,

$$y^{(2)}(x) = y_0 + \int_0^x f(x, y^{(1)}(x))dx$$

$$= 1 + \int_0^x \frac{x - 2x + 2\log_e|1+x|}{1+2\log_e|1+x|}dx = 1 + x - 2\int_0^x \frac{x}{1+2\log_e|1+x|}dx$$

We observe that, it is not possible to obtain the integral for getting $y^{(2)}(x)$. Thus Picard's method is not applicable to get successive approximate solutions.

### Multistep Methods

We have seen that for finding the solution at each step, the Taylor series method and Runge-Kutta methods requires evaluation of several derivatives. We shall

NOTES

now develop the multistep method which require only one derivative evaluation per step; but unlike the self starting Taylor series or Runge-Kutta methods, the multistep methods make use of the solution at more than one previous step points.

Let the values of $y$ and $y^1$ already have been evaluated by self-starting methods at a number of equally spaced points $x_0, x_1,..., x_n$. We now integrate the differential equation,

$$\frac{dy}{dx} = f(x, \ y), \text{ from } x_n \text{ to } x_{n+1}$$

*i.e.,*
$$\int_{x_n}^{x_{n+1}} dy = \int_{x_n}^{x_{n+1}} f(x, y)dx$$

$\therefore$
$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x, y(x))\, dx$$

To evaluate the integral on the right hand side, we consider $f(x, y)$ as a function of $x$ and replace it by an interpolating polynomial, i.e., a Newton's backward difference interpolation using the $(m + 1)$ points $x_n, x_{n+1}, x_{n-2},..., x_{n-m}$,

$$p_m(x) = \sum_{k=0}^{m} (-1)^k \binom{-s}{k} \Delta^k f_{n-k}, \quad \text{where} \quad s = \frac{x - x_n}{h}$$

$$\binom{-s}{k} = -s(-s-1)(-s-2)...(-s-k+1) \cdot \frac{1}{k!}$$

Substituting $p_m(x)$ in place of $f(x, y)$, we obtain

$$y_{n+1} = y_n + h\int_0^1 \sum_{k=0}^{m} (-1)^k \ \binom{-s}{k} \ \Delta^k f_{n-k} \ ds$$

$$= y_n + h \ \ [\gamma f_n + \gamma_1 \Delta f_{n-1} + \gamma_2 \Delta^2 f_{n-2} + ... + \gamma_m \Delta f_{n-m}]$$

Where $\gamma_k = (-1)^k \int_0^1 \binom{-s}{k} ds$

The coefficients $\gamma_k$ can be easily computed to give,

$$\gamma_0 = 1, \ \gamma_1 = \frac{1}{2}, \ \gamma_2 = \frac{5}{12}, \ \gamma_3 = \frac{3}{8}, \ \gamma_4 = \frac{251}{720}, \text{ etc.}$$

Taking $m = 3$, the above formula gives,

$$y_{n+1} = y_n + h\left[ f_n + \frac{1}{2}\Delta f_{n-1} + \frac{5}{12}\Delta^2 f_{n-2} + \frac{3}{8}\Delta^3 f_{n-3} \right]$$

Substituting the expression of the differences in terms of function values given by,

$$\Delta f_{n-1} = f_n - f_{n-1}, \ \Delta^2 f_{n-2} = f_n - 2f_{n-1} + f_{n-2}$$
$$\Delta^3 f_{n-3} = f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}$$

NOTES

We get on arranging,

$$y_{n+1} = y_n + \frac{h}{24}[55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}] \qquad (10.8)$$

This is known as *Adams-Bashforth formula of order* 4. The local error of this formula is,

$$E = h^5 f^{iv}(\xi)\int_0^1 \binom{s+3}{4} ds \qquad (10.9)$$

By using mean value theorem of integral calculus,

$$E = h^5 f^{iv}(\eta)\int_0^1 \binom{s+3}{4} ds$$

Or,

$$E = h^5 f^{iv}(\eta).\frac{251}{720} \qquad (10.10)$$

The fourth order Adams-Bashforth formula requires four starting values, i.e., the derivaties, $f_3, f_2, f_1$ and $f_0$. This is a multistep method.

**Predictor-Correction Methods**

These methods use a pair of multistep numerical integration. The first is the Predictor formula, which is an open-type explicit formula derived by using, in the integral, an interpolation formula which interpolates at the points $x_n, x_{n-1}, ..., x_{n-m}$. The second is the Corrector formula which is obtained by using interpolation formula that interpolates at the points $x_{n+1}, x_n, ..., x_{n-p}$ in the integral.

**Euler's Predictor-Corrector Formula**

The simplest formula of the type is a pair of formula given by,

$$y_{n+1}^{(p)} = y_n + h\,f(x_n, y_n) \qquad (10.11)$$

$$y_{n+1}^{(c)} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(p)})] \qquad (10.12)$$

In order to determine the solution of the problem upto a desired accuracy, the corrector formula can be employed in an iterative manner as shown below:

**Step 1:** Compute $y_{n+1}^{(0)}$, using Equation (10.11)

i.e., $y_{n+1}^{(0)} = y_n + h\,f(x_n, y_n)$

**Step 2:** Compute $y_{n+1}^{(k)}$ using Equation (10.12)

i.e., $y_{n+1}^{(k)} = y_n + \frac{h}{2}\ [f(x_n,\ \ y_n) + f(x_{n+1},\ \ y_{n+1}^{(k-1)})]$, for $K = 1, \ 2, \ 3, ...,$

The computation is continued till the condition given below is satisfied,

$$\left|\frac{y_{n+1}^{(k)} - y_{n+1}^{(k-1)}}{y_{n+1}^{(k)}}\right| < \varepsilon \qquad (10.13)$$

NOTES

where ε is the prescribed accuracy.

It may be noted that the accuracy achieved will depend on step size $h$ and on the local error. The local error in the predictor and corrector formula are,

$$\frac{h^2}{2}y''(\eta_1) \quad \text{and} \quad -\frac{h^3}{12}y'''(\eta_2), \text{ respectively.}$$

**Milne's Predictor-Corrector Formula**

A commonly used Predictor-Corrector system is the fourth order *Milne's Predictor-Corrector* formula. It uses the following as Predictor and Corrector.

$$y_{n+1}^{(p)} = y_{n-3} + \frac{4h}{3}(2f_n - f_{n-1} + 2f_{n-2}^{*})$$

$$y_{n+1}^{(c)} = y_{n-1} + \frac{h}{3}[f_{n-1} + 4f_n + f_{n+1}(x_{n+1}, y_{n+1}^{(p)})] \qquad (10.14)$$

The local errors in these formulae are respectively,

$$\frac{14}{45}h^5 y^{(v)}(\xi_1) \text{ and } -\frac{1}{90}h^5 y^{(v)}(\xi_2) \qquad (10.15)$$

**Example 5:** Compute the Taylor series solution of the problem $\frac{dy}{dx} = xy + 1,\ y(0) = 1,$ up to $x^5$ terms and hence compute values of $y(0.1)$, $y$ (0.2) and $y$ (0.3). Use Milne's Predictor-Corrector method to compute $y$ (0.4) and $y$ (0.5).

**Solution:** We have $y' = xy + 1$, with $y$ (0) = 1, $\therefore y'(0) = 1$

Differentiating successively, we get

$$y''(x) = xy' + y \qquad \therefore y''(0) = 1$$
$$y'''(x) = xy'' + y' \qquad \therefore y'''(0) = 2$$
$$y^{(iv)}(x) = xy''' + 3y'' \qquad \therefore y^{(iv)}(0) = 3$$
$$y^{(v)}(x) = xy^{(iv)} + 4y''' \qquad \therefore y^{(iv)}(0) = 8$$

Thus the Taylor series solution is given by,

$$y(x) = y(0) + xy'(0) + \frac{x^2}{2}y''(0) + \frac{x^3}{3!}y'''(0) + \frac{x^4}{4!}y^{iv}(0) + \frac{x^5}{5!}y^{(v)}(0)$$
$$= 1 + x + \frac{x^2}{2} + \frac{x^3}{3!}\times 2 + \frac{x^4}{4!}\times 3 + \frac{x^5}{5!}\times 8$$
$$\therefore y(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{8} + \frac{x^5}{15}$$
$$\therefore \quad y(0.1) = 1 + 0.1 + \frac{0.01}{2} + \frac{0.001}{3} + \frac{0.0001}{8} + \frac{0.00001}{15}$$
$$= 1.1053$$
$$y(0.2) = 1 + 0.2 + \frac{0.04}{2} + \frac{0.008}{3} + \frac{0.0016}{8} + \frac{0.00032}{15}$$
$$= 1.22288$$
$$y(0.3) = 1 + 0.3 + \frac{0.09}{2} + \frac{0.027}{3} + \frac{0.0081}{8} + \frac{0.00243}{15}$$
$$= 1.35526$$

NOTES

For application of Milne's Predictor-Corrector method, we compute $y'(0.1)$, $y'(0.2)$ and $y'(0.3)$.

$$y'(0.1) = 0.1 \times 1.1053 + 1 = 1.11053$$
$$y'(0.2) = 0.2 \times 1.22288 + 1 = 1.244576$$
$$y'(0.3) = 0.3 \times 1.35526 + 1 = 1.40658$$

The Predictor formula gives, $y_4 = y(0.4) = y_0 + \frac{4h}{3}(2y_1' - y_2' + 2y_3')$.

$$\therefore y_4^{(0)} = 1 + \frac{4 \times 0.1}{3}(2 \times 1.11053 - 1.24458 + 2 \times 1.40658)$$
$$= 1.50528 \qquad \therefore y_4' = 1 + 0.4 \times 1.50528 = 1.602112$$

The Corrector formula gives, $y_4^{(1)} = y_2 + \frac{h}{3}(y_2' + 4y_3' + y_4')$.

$$y(0.4) = 1.22288 + \frac{0.1}{3}(1.24458 + 4 \times 1.40658 + 1.60211)$$
$$= 1.22288 + 0.28243$$
$$= 1.50531$$

**Numerical Solution of Boundary Value Problems**

We consider the solution of ordinary differential equation of order 2 or more, when value of the dependent variable is given at more than one point, usually at the two ends of an interval in which the solution is required. For example, the simplest boundary value problem associated with a second order differential equation is,

$$y'' + p(x)\,y' + q(x)y = r(x) \qquad (10.16)$$

with boundary conditions, $y(a) = A, y(b) = B$. (10.17)

The following two methods reduce the boundary value problem into initial value problems which are then solved by any of the methods for solving such problems.

**Reduction to a Pair of Initial Value Problem**

This method is applicable to linear differential equations only. In this method, the solution is assumed to be a linear combination of two solutions in the form,

$$y(x) = u(x) + \lambda \upsilon(x) \qquad (10.18)$$

where $\lambda$ is a suitable constant determined by using the boundary condition and $u(x)$ and $\upsilon(x)$ are the solutions of the following two initial value problems:

(*i*) $u'' + p(x)u' + q(x)u = r(x)$

$u(a) = A,\ u'(a) = \alpha_1$, (say). (10.19)

(*ii*) $\upsilon'' + p(x)\upsilon' + q(x)\upsilon = r(x)$

$\upsilon(a) = 0$ and $\upsilon'(a) = \alpha_2$, (say) (10.20)

NOTES

where $\alpha_1$ and $\alpha_2$ are arbitrarily assumed constants. After solving the two initial value problems, the constant $\lambda$ is determined by satisfying the boundary condition at $x = b$. Thus,

$$B = u(b) + \lambda \upsilon (b)$$

Or, $$\lambda = \frac{B - \upsilon (b)}{\upsilon (b)}, \text{ provided } \upsilon (b) \neq 0 \quad (10.21)$$

Evidently, $y(a) = A$, is already satisfied.

If $\upsilon(b) = 0$, then we solve the initial value problem for $\upsilon$ again by choosing $\upsilon'(a) = \alpha_3$, for some other value for which $\upsilon(b)$ will be non-zero.

Another method which is commonly used for solving boundary problems is the finite difference method discussed below.

### Finite Difference Method

In this method of solving boundary value problem, the derivatives appearing in the differential equation and boundary conditions, if necessary, are replaced by appropriate difference gradients.

Consider the differential equation, $y'' + p(x)\, y' + q(x)y = r(x)$ (10.22)

with the boundary conditions, $y\,(a) = \alpha$ and $y\,(b) = \beta$ (10.23)

The interval $[a, b]$ is divided into $N$ equal parts each of width $h$, so that $h = (b–a)/N$, and the end points are $x_0 = a$ and $x_n = b$. The interior mesh points $x_i$ at which solution values $y(x_i)$ are to be determined are,

$$x_n = x_0 + nh,\ n = 1, 2, ..., N - 1 \quad (10.24)$$

The values of $y$ at the mesh points is denoted by $y_n$ given by,

$$y_n = y\,(x_0 + nh),\ n = 0, 1, 2, ..., N \quad (10.25)$$

The following central difference approximations are usually used in finite difference method of solving boundary value problem,

$$y'(x_n) \approx \frac{y_{n+1} - y_{n-1}}{2h} \quad (10.26)$$

$$y''(x_n) \approx \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} \quad (10.27)$$

Substituting these in the differential equation, we have

$2(y_{n+1} – 2y_n + y_{n–1}) + p_n\, h(y_{n+1} – y_{n–1}) + 2h^2 g_n y_n = 2r_n h^2$,

where $p_n = p(x_n)$, $q_n = q(x_n)$, $r_n = r(x_n)$ (10.28)

Rewriting the equation by regrouping we get,

$$(2–hp_n)y_{n–1} + (–4+2h^2 q_n)y_n + (2+h^2 q_n)y_{n+1} = 2r_n h^2 \quad (10.29)$$

NOTES

This equation is to be considered at each of the interior points, i.e., it is true for $n = 1, 2, ..., N–1$.

The boundary conditions of the problem are given by,

$$y_0 = \alpha, \quad y_n = \beta \tag{10.30}$$

Introducing these conditions in the relevant equations and arranging them, we have the following system of linear equations in $(N–1)$ unknowns $y_1, y_2, ..., y_{n-1}$.

$$\begin{aligned}
&(-4+2h^2q_1)y_1 + (2+hp_1)y_2 = 2r_1h^2 - (2-hp_1)\alpha \\
&(2-hp_2)y_1 + (-4+2h^2q_2)y_2 + (2+hp_2)y_3 = 2r_2h^2 \\
&(2-hp_3)y_2 + (-4+2h^2q_3)y_3 + (2+hp_3)y_4 = 2r_3h^2 \\
&\dots \quad \dots \quad \dots \quad \dots \quad \dots \\
&(2-hp_{N-2}) + (-4+2h^2q_{N-2})y_{N-2} \quad (2+hp_{N-2})y_{N-1} = 2r_{N-2}h^2 \\
&(2-hp_{N-1}) \quad y_{N-2} + (-4+2h^2q_{N-1})y_{N-1} = 2r_{N-1}h^2 - (2-hp_{N-1})\beta
\end{aligned} \tag{10.31}$$

The above system of $N–1$ equations can be expressed in matrix notation in the form

$$Ay = b \tag{10.32}$$

Where the coefficient matrix $A$ is a tridiagonal one, of the form

$$A = \begin{bmatrix} B_1 & C_1 & 0 & 0... & 0 & 0 & 0 \\ A_2 & B_2 & C_2 & 0... & 0 & 0 & 0 \\ 0 & A_3 & B_3 & C_3... & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0... & A_{N-2} & B_{N-2} & C_{N-2} \\ 0 & 0 & 0 & 0... & 0 & A_{N-1} & B_{N-1} \end{bmatrix} \tag{10.33}$$

Where

$$\begin{aligned} B_i &= -4 + 2h^2 q_i, \quad i = 1, 2, ..., N-1 \\ C_i &= 2 + hp_i, \qquad i = 1, 2, ..., N-2 \\ A_i &= 2 - hp_i, \qquad i = 2, 3, ..., N-1 \end{aligned} \tag{10.34}$$

The vector $b$ has components,

$$\begin{aligned} b_1 &= 2\gamma_1 h^2 - (2 - hp_1)\alpha \\ b_i &= 2\gamma_i h^2, \text{for } i = 2, 3, ..., N-2 \\ b_{N-1} &= 2\gamma_{N-1} + h^2 - (2 - hlp_{N-1})\beta \end{aligned} \tag{10.35}$$

The system of linear equations can be directly solved using suitable methods.

NOTES

**Example 6:** Compute values of $y$ (1.1) and $y$ (1.2) on solving the following initial value problem, using Runge-Kutta method of order 4:

$$y'' + \frac{y'}{x} + y = 0, \text{ with } y(1) = 0.77,\ y'(1) = -0.44$$

**Solution:** We first rewrite the initial value problem in the form of pair of first order equations.

$$y' = z,\ z' = \frac{-z}{x} - y$$

with $y$ (1) = 0.77 and $z$ (1) = –0.44.

We now employ Runge-Kutta method of order 4 with $h$ = 0.1,

$$y(1.1) = y(1) + \frac{1}{6}\ (k_1 + 2k_2 + 2k_3 + k_4)$$

$$y'(1.1) = z(1.1) = 1 + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)$$

$$k_1 = -0.44 \times 0.1 = -0.044$$

$$l_1 = 0.1 \times \left(\frac{0.44}{1} - 0.77\right) = -0.033$$

$$k_2 = 0.1 \times \left(-0.44 - \frac{0.033}{2}\right) = 0.04565$$

$$l_2 = 0.1 \times \left(\frac{0.4565}{1.05} - 0.748\right) = -0.031323809$$

$$k_3 = 0.1 \times \left(-0.44 + \frac{-0.03123809}{2}\right) = -0.0455661904$$

$$l_3 = 0.1 \times \left[\frac{0.0455661904}{1.05} - 0.747175\right] = -0.031321128$$

$$k_4 = 0.1 \times (-0.47132112) = -0.047132112$$

$$l_4 = 0.1 \times \left(\frac{0.047132112}{1.1} - 0.72443381\right) = -0.068158643$$

$$\therefore\ y(1.1) = 0.77 + \frac{1}{6}[-0.044 + 2 \times (-0.045661904) - 0.029596005] = 0.727328602$$

$$y'(1.1) = -0.44 + \frac{1}{6}\ [-0.033 + 2(-0.031323809) + 2(-0.031321128) - 0.029596005]$$

$$= -0.44 + \frac{1}{6}[-0.33 - 0.062647618 - 0.062642256 - 0.029596005]$$

$$= -0.526322021$$

**Example 7:** Compute the solution of the following initial value problem for $x$ = 0.2, using Taylor series solution method of order 4: n.l.

$$\frac{d^2y}{dx^2} = y + x\frac{dy}{dx},\quad y(0) = 1,\quad y'(0) = 0$$

NOTES

**Solution:** Given $y'' = y + xy'$, we put $z = y'$, so that

$z' = y + xz,\ y' = z$ and $y(0) = 1,\ z(0) = 0.$

We solve for $y$ and $z$ by Taylor series method of order 4. For this we first compute $y''(0),\ y'''(0),\ y^{iv}(0),...$

*We have,*
$$y''(0) = y(0) + 0 \times y'(0) = 1, \quad z'(0) = 1$$
$$y'''(0) = z''(0) = y'(0) + z(0) + 0.z'(0) = 0$$
$$y^{iv}(0) = z'''(0) = y''(0) + 2z'(0) + 0.z''(0) = 3$$
$$z^{iv}(0) = 4z''(0) + 0.z'''(0) = 0$$

By Taylor series of order 4, we have

$$y(0+x) = y(0) + xy'(0) + \frac{x^2}{2!}y''(0) + \frac{x^3}{3!}y'''(0) + \frac{x^4}{4!}y^{iv}(0)$$

*or,*
$$y(x) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} \times 3$$

$$\therefore y(0.2) = 1 + \frac{(0.2)^2}{2!} + \frac{(0.2)^4}{8} = 1.0202$$

Similarly, $y'(0.2) = z(0.2) = 0.2 + \dfrac{(0.2)^3}{4!} \times 3 = 0.204$

**Example 8:** Compute the solution of the following initial value problem for $x$ = 0.2 by fourth order Runge -Kutta method: n.l. $\dfrac{d^2y}{dx^2} = xy,\ y(0) = 1,\ y'(0) = 1$

**Solution:** Given $y'' = xy$, we put $y' = z$ and the simultaneous first order problem,

$y' = z = f(x, y, z)$, say $z' = xy = g(x, y, z)$, say with $y(0) = 1$ and $z(0) = 1$

We use Runge-Kutta 4th order formulae, with $h = 0.2$, to compute $y$ (0.2) and $y'(0.2)$, given below.

$$k_1 = h\ f(x_0,\ \ y_0,\ \ z_0) = 0.2 \times 1 = 0.2$$
$$l_1 = h\ g(x_0,\ \ y_0,\ \ z_0) = 0.2 \times 0 = 0$$
$$k_2 = h\ f\left(x_0 + \frac{h}{2},\ \ y_0 + \frac{k_1}{2},\ \ z_0 + \frac{l_1}{2}\right) = 0.2 \times (1 + 0) = 0.2$$
$$l_2 = h\ g\left(x_0 + \frac{h}{2},\ \ y_0 + \frac{k_1}{2},\ \ z_0 + \frac{l_1}{2}\right) = 0.2 \times \frac{0.2}{2}\ \left(1 + \frac{0.2}{2}\right) = 0.022$$
$$k_3 = h\ f\left(x_0 + \frac{h}{2},\ \ y_0 + \frac{k_2}{2}, z_0 + \frac{l_2}{2}\right) = 0.2 \times 1.011 = 0.2022$$
$$l_3 = h\ g\left(x_0 + \frac{h}{2},\ \ y_0 + \frac{k_2}{2},\ \ z_0 + \frac{l_2}{2}\right) = 0.2 \times 0.1 \times 1.1 = 0.022$$
$$k_4 = h\ f(x_0 + h,\ \ y_0 + k_3,\ \ z_0 + l_3) = 0.2 \times 1.022 = 0.2044$$
$$l_4 = h\ \ g(x_0 + h,\ \ y_0 + k_3,\ \ z_0 + l_3) = 0.2 \times 0.2 \times 1.2022 = 0.048088$$
$$y(0.2) = 1 + \frac{1}{6}\ \ (0.2 + 2(0.2 + 0.2022) + 0.2044) = 1.2015$$
$$y'(0.2) = 1 + \frac{1}{6}\ \ (0 + 2\ \ (0.022 + 0.022) + 0.048088) = 1.02268$$

NOTES

### Local Truncation Error

Local Truncation error in a numerical method is error that is caused by using simple approximations to represent exact mathematical formulas. The only way to completely avoid truncation error is to use exact calculations. However, truncation error can be reduced by applying the same approximation to a larger number of smaller intervals or by switching to a better approximation. Analysis of truncation error is the single most important source of information about the theoretical characteristics that distinguish better methods from poorer ones. With a combination of theoretical analysis and numerical experiments, it is possible to estimate truncation error accurately.

**Check Your Progress**

1. Define Picard's method of successive approximation.
2. What is a predictor formula?
3. What are local errors in Milne's predictor-corrector formulae?
4. Where can the method of reduction to a pair of initial value problem be applied?

## 10.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In Picard's method the first approximate solution $y^{(1)}(x)$ is obtained by replacing $y(x)$ by $y_0$. Thus, $y^{(1)}(x) = y_0 + \int_{x_0}^{x} f(x, y_0)dx$. The second approximate solution is derived on replacing $y$ by $y(1)$ $(x)$. Thus,

$$y^{(2)}(x) = y_0 + \int_{x_0}^{x} f(x, y^{(1)}(x))\, dx.$$

   This iteration formula is known as Picard's iteration for finding solution of a first order differential equation, when an initial condition is given. The iterations are continued until two successive approximate solutions $y^k$ and $y^{k+1}$ give approximately the same result for the desired values of $x$ up to a desired accuracy.

2. A predictor formula is an open-type explicit formula derived by using, in the integral, an interpolation formula which interpolates at the points $x_n, x_{n-1}, \ldots, x_{n-m}$.

3. The local errors in these formulae are $\frac{14}{45}h^5 y^{(v)}(\xi_1)$ and $-\frac{1}{90}h^5 y^{(v)}(\xi_2)$.
4. This method is applicable to linear differential equations only.

NOTES

## 10.4 SUMMARY

- Picard's iteration is a method of finding solutions of a first order differential equation when an initial condition is given.
- The multistep method requires only one derivative evaluation per step; but unlike the self starting Taylor series on Runge-Kutta methods, the multistep methods make use of the solution at more than one previous step points.
- These methods use a pair of multistep numerical integration. The first is the predictor formula, which is an open-type explicit formula derived by using, in the integral, an interpolation formula which interpolates at the points $x_n, x_{n-1}, ..., x_{n-m}$. The second is the corrector formula which is obtained by using interpolation formula that interpolates at the points $x_{n+1}, x_n, ..., x_{n-p}$ in the integral.
- The solution of ordinary differential equation of order 2 or more, when values of the dependent variable is given at more than one point, usually at the two ends of an interval in which the solution is required.
- The methods used to reduce the boundary value problem into initial value problems are reduction to a pair of initial value problem and finite difference method.

## 10.5 KEY WORDS

- **Predictor formula:** It is an open-type explicit formula derived by using, in the integral, an interpolation formula which interpolates at the points $x_n$, $x_{n-1}$, ..., $x_{n-m}$.
- **Corrector formula:** It is obtained by using interpolation formula that interpolates at the points $x_{n+1}, x_n$, ..., $x_{n-p}$ in the integral.

## 10.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. What are ordinary differential equations?
2. Name the methods for computing the numerical solution of differential equations.

3. When is multistep method used?
4. Name the predictor-corrector methods.
5. How will you find the numerical solution of boundary value problems?

NOTES

**Long-Answer Questions**

1. Use Picard's method to compute values of $y(0.1), y(0.2)$ and $y(0.3)$ correct to four decimal places, for the problem, $y' = x + y, \quad y(0) = 1$.
2. Given $\frac{dy}{dx} = \frac{1}{2}(1 + x^2)y^2$, and $y(0) = 1, y(0.1) = 1.06, y(0.2) = 1.12, y(0.3) = 1.21$. Compute $y(0.4)$ by Milne's predictor-corrector method.

## 10.7 FURTHER READINGS


Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

# UNIT 11 EULER'S METHOD

NOTES

**Structure**



## 11.0 INTRODUCTION

The Euler method is a first-order method, which means that the local error (error per step) is proportional to the square of the step size, and the global error (error at a given time) is proportional to the step size. The Euler method often serves as the basis to construct more complex methods.

In this unit, you will study about the Euler's method and modified Euler's method.

## 11.1 OBJECTIVES

After going through this unit, you will be able to:

- Analyse the Euler's method
- Understand about the modified Euler's method

## 11.2 EULER'S METHOD

Euler's is a crude but simple method of solving a first order initial value problem:

$$\frac{dy}{dx} = f(x, y),\ y(x_0) = y_0$$

This is derived by integrating $f(x_0, y_0)$ instead of $f(x, y)$ for a small interval,

$$\therefore \quad \int_{x_0}^{x_0+h} dy = \int_{x_0}^{x_0+h} f(x_0, y_0)dx$$

$$\therefore \quad y(x_0 + h) - y(x_0) = hf(x_0, y_0)$$

Writing $y_1 = y(x_0 + h)$, we have

$$y_1 = y_0 + h f(x_0, y_0) \qquad (11.1)$$

Similarly, we can write

$$y_2 = y\,(x_1 + h) = y_1 + h\,f\,(x_1, y_1) \qquad (11.2)$$

where $x_1 = x_0 + h$.

Proceeding successively, we can get the solution at any $x_n = x_0 + nh$, as

$$y_n = y_{n-1} + h\,f\,(x_{n-1}, y_{n-1}) \qquad (11.3)$$

This method, known as Euler's method, can be geometrically interpreted, as shown in Figure 11.1.

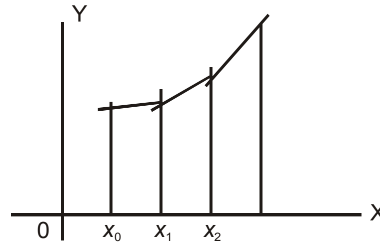



***Fig. 11.1*** *Euler's Method*

For small step size $h$, the solution curve $y = y\,(x)$, is approximated by the tangential line.

The local error at any $x_k$, i.e., the truncation error of the Euler's method is given by,

$$e_k = y(x_{k+1}) - y_{k+1}$$

Where $y_{k+1}$ is the solution by Euler's method.

$$\therefore \quad e_k = y(x_k + h) - \{y_k + hf(x_k, y_k)\}$$

$$= y_k + hy'(x_k) + \frac{h^2}{2}y''(x_k + \theta h) - y_k - hy'(x_k), \;\; 0 < \theta < 1$$

$$\therefore \quad e_k = \frac{h^2}{2}y''(x_k + \theta h), \;\; 0 < \theta < 1$$

***Note:*** The Euler's method finds a sequence of values $\{y_k\}$ of $y$ for the sequence of values $\{x_k\}$ of $x$, step by step. But to get the solution up to a desired accuracy, we have to take the step size $h$ to be very small. Again, the method should not be used for a larger range of $x$ about $x_0$, since the propagated error grows as integration proceeds.

**Example 1:** Solve the following differential equation by Euler's method for $x = 0.1$, 0.2, 0.3; taking $h = 0.1$; $\frac{dy}{dx} = x^2 - y,\; y(0) = 1$. Compare the results with exact solution.

**Solution:** Given $\frac{dy}{dx} = x^2 - y$, with $y\,(0) = 1$.

In Euler's method one computes in successive steps, values of $y_1, y_2, y_3,...$ at $x_1 = x_0 + h, x_2 = x_0 + 2h, x_3 = x_0 + 3h$, using the formula,

$$y_{n+1} = y_n + hf(x_n, \ y_n), \text{ for } n = 0, \ 1, \ 2,...$$

$$\therefore \qquad y_{n+1} = y_n + h \ \ (x^2{}_n - y_n)$$

With $h$ = 0.1 and starting with $x_0$ = 0, $y_0$ = 1, we present the successive computations in the table given below.

| $n$ | $x_n$ | $y_n$ | $f(x_n, y_n) = x^2{}_n - y_n$ | $y_{n+1} = y_n + hf(x_n, y_n)$ |
|---|---|---|---|---|
| 0 | 0.0 | 1.000 | −1.000 | 0.9000 |
| 1 | 0.1 | 0.900 | −0.8900 | 0.8110 |
| 2 | 0.2 | 0.8110 | −0.7710 | 0.7339 |
| 3 | 0.3 | 0.7339 | −0.6439 | 0.6695 |

The analytical solution of the differential equation written as $\frac{dy}{dx} + y = x^2$, is

$$ye^x = \int x^2 e^x dx + c$$

Or, $$ye^x = x^2e^x - 2xe^x + 2e^x + c.$$

Since, $$y = 1 \text{ for } x = 0, \therefore c = -1.$$

$$\therefore \qquad y = x^2 - 2x + 2 - e^{-x}.$$

The following table compares the exact solution with the approximate solution by Euler's method.

| $n$ | $x_n$ | Approximate Solution | Exact Solution | % Error |
|---|---|---|---|---|
| 1 | 0.1 | 0.9000 | 0.9052 | 0.57 |
| 2 | 0.2 | 0.8110 | 0.8213 | 1.25 |
| 3 | 0.3 | 0.7339 | 0.7492 | 2.04 |

**Example 2:** Compute the solution of the following initial value problem by Euler's method, for $x$ = 0.1 correct to four decimal places, taking $h$ = 0.02, $\frac{dy}{dx} = \frac{y-x}{y+x}, y(0) = 1$.

**Solution:** Euler's method for solving an initial value problem,

$$\frac{dy}{dx} = f(x, y), \ y(x_0) = y_0, \text{ is } y_{n+1} = y_n + h\ f(x_n, y_n), \text{ for } n = 0, 1, 2,...$$

Taking $h = 0.02$, we have $x_1 = 0.02, x_2 = 0.04, x_3 = 0.06, x_4 = 0.08, x_5 = 0.1$.

Using Euler's method, we have, since $y(0) = 1$

$$y(0.02) = y_1 = y_0 + h\,f(x_0, y_0) = 1 + 0.02 \times \frac{1-0}{1+0} = 1.0200$$

$$y(0.04) = y_2 = y_1 + h\,f(x_1, y_1) = 1.0200 + 0.02 \times \frac{1.0200-0.02}{1.0200+0.02} = 1.0392$$

NOTES

$$y(0.06) = y_3 = y_2 + h\,f(x_2, y_2) = 1.0392 + 0.02 \times \frac{1.0392-0.04}{1.0392+0.04} = 1.0577$$

$$y(0.08) = y_4 = y_3 + h\,f(x_3, y_3) = 1.0577 + 0.02 \times \frac{1.0577-0.06}{1.0577+0.06} = 1.0756$$

$$y(0.1) = y_5 = y_4 + h\,f(x_4, y_4) = 1.0756 + 0.02 \times \frac{1.0756-0.08}{1.0756+0.08} = 1.0928$$

Hence, $y\,(0.1) = 1.0928$.

### Modified Euler's Method

In order to get somewhat moderate accuracy, Euler's method is modified by computing the derivative $y' = f(x, y)$, at a point $x_n$ as the mean of $f(x_n, y_n)$ and $f(x_{n+1}, y^{(0)}{}_{n+1})$, where,

$$y^{(0)}{}_{n+1} = y_n + h \quad f(x_n,\ y_n)$$

$$y^{(1)}{}_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1},\ y^{(0)}{}_{n+1})] \qquad (11.4)$$

This modified method is known as Euler-Cauchy method. The local truncation error of the modified Euler's method is of the order $O(h^3)$.

***Note:*** Modified Euler's method can be used to compute the solution up to a desired accuracy by applying it in an iterative scheme as stated below.

$$\text{Compute } y^{(k)}{}_{n+1} = y_n + h\ f(x_n,\ y_n)$$

$$\text{Compute } y_{n+1}^{(k+1)} = y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(k)})\right], \text{ for } k = 0, 1, 2, \ldots \qquad (11.5)$$

The iterations are continued until two successive approximations $y_{n+1}^{(k)}$ and $y_{n+1}^{(k+1)}$ coincide to the desired accuarcy. As a rule, the iterations converge rapidly for a sufficiently small $h$. If, however, after three or four iteration the iterations still do not give the necessary accuracy in the solution, the spacing $h$ is decreased and iterations are performed again.

**Example 3:** Use modified Euler's method to compute $y\,(0.02)$ for the initial value problem, $\frac{dy}{dx} = x^2 + y,$ with $y\,(0) = 1$, taking $h = 0.01$. Compare the result with the exact solution.

**Solution:** Modified Euler's method consists of obtaining the solution at successive points, $x_1 = x_0 + h, x_2 = x_0 + 2h, \ldots, x_n = x_0 + nh$, by the two stage computations given by,

$$y_{n+1}^{(0)} = y_n + hf(x_n, y_n)$$

$$y_{n+1}^{(1)} = y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(0)})\right]$$

NOTES

For the given problem, $f(x, y) = x^2 + y$ and $h = 0.01$

$$y_1^{(0)} = y_0 + h[x_0^2 + y_0] = 1 + 0.01 \times 1 = 1.01$$

$$y_1^{(1)} = 1 + \frac{0.01}{2}[1.0 + 1.01 + (0.01)^2] = 1.01005$$

i.e., $y_1 = y \quad (0.01) = 1.01005$

Next, $y_2^{(0)} = y_1 + h[x_1^2 + y_1]$

$$= 1.01005 + 0.01[(0.1)^2 + 1.01005]$$

$$= 1.01005 + 0.010102 = 1.02015$$

$$y_2^{(1)} = 1.01005 + \frac{0.01}{2}[(0.01)^2 + 1.01005 + (0.01)^2 + 1.02015]$$

$$= 1.01005 + \frac{0.01}{2} \times \ (2.02140)$$

$$= 1.01005 + 0.10107$$

$$= 1.11112$$

$$\therefore y_2 = y \quad (0.02) = 1.11112$$

### Euler's Method for a Pair of Differential Equations

Consider an initial value problem associated with a pair of first order differential equation given by,

$$\frac{dy}{dx} = f(x, y, z), \frac{dz}{dx} = g(x, y, z) \qquad (11.6)$$

$$\text{with } y(x_0) = y_0, z(x_0) = z_0 \qquad (11.7)$$

Euler's method can be extended to compute approximate values $y_i$ and $z_i$ of $y(x_i)$ and $z(x_i)$ respectively given by,

$$y_{i+1} = y_i + h f(x_i, y_i, z_i)$$
$$z_{i+1} = z_i + h g(x_i, y_i, z_i) \qquad (11.8)$$

starting with $i = 0$ and continuing step by step for $i = 1, 2, 3,...$ Evidently, we can also extend Euler's method for an initial value problem associated with a second order differential equation by rewriting it as a pair of first order equations.

Consider the initial value problem,

$$\frac{d^2y}{dx^2} = g\left(x, y, \frac{dy}{dx}\right), \text{ with } y(x_0) = y_0,\ y'(x_0) = y'_0$$

We write $\frac{dy}{dx} = z,$ so that $\frac{dz}{dx} = g(x, y, z)$ with $y(x_0) = y_0$ and $z(x_0) = y'_0$.

**Example 4:** Compute $y(1.1)$ and $y(1.2)$ by solving the initial value problem,

$$y'' + \frac{y'}{x} + y = 0, \text{ with } y(1) = 0.77,\ y'(1) = -0.44$$

**Solution:** We can rewrite the problem as $y' = z,\ z' = -\frac{z}{x} - y$; with $y$, (1) = 0.77 and $z$ (1.1) = –0.44.

Taking $h = 0.1$, we use Euler's method for the problem in the form,

NOTES

$$y_{i+1} = y_i + hz_i$$

$$z_{i+1} = z_i + h\left[-\frac{z_1}{x_i} - y_i\right],\ i = 0, 1, 2, ...$$

Thus, $y_1 = y$ (1.1) and $z_1 = z$ (1.1) are given by,

$$y_1 = y_0 + hz_0 = 0.77 + 0.1 \times (-0.44) = 0.726$$

$$z_1 = z_0 + h\left[-\frac{z_0}{x_0} - y_0\right] = -0.44 + 0.1 \times (0.44 - 0.77)$$

$$= -0.44 - 0.33 = -0.473$$

Similarly, $y_2 = y(1.2) = y_1 + hz_1 = 0.726 - 0.1\ (-0.473) = 0.679$

$$z_2 = z(1.2) = z_1 + h\ \left[-\frac{z_1}{x_1} - y_1\right]$$

$$= -0.473 + 0.1 \times \left(\frac{0.473}{1.1} - 0.726\right)$$

$$= -0.473 + 0.1 \times -0.296 = -0.503$$

Thus, $y$ (1.1) = 0.726 and $y$ (12) = 0.679.

**Example 5:** Using Euler's method, compute $y$ (0.1) and $y$ (0.2) for the initial value problem,

$$y'' + y = 0,\ y(0) = 0,\ y'(0) = 1$$

**Solution:** We rewrite the initial value problem as $y' = z,\ z' = -y$, with $y$ (0) = 0, $z$ (0) = 1.

Taking $h = 0.1$, we have by Euler's method,

$$\begin{aligned} y_1 &= y(0.1) = y_0 + hz_0 && = 0 + 0.1 \times 1 && = 0.1 \\ z_1 &= z(0.1) = z_0 + h(-y_0) && = 1 + 0.1 \times 0 && = 1.0 \\ y_2 &= y(0.2) = y_1 + hz_1 && = 0.1 + 0.1 \times 1.0 && = 0.2 \\ z_2 &= z(0.2) = z_1 - hy_1 && = 1.0 - 0.1 \times 0.1 && = 0.99 \end{aligned}$$

**Example 6:** For the initial value problem $y'' + xy' + y = 0$, $y(0) = 0$, $y'$ (0) = 1. Compute the values of $y$ for 0.05, 0.10, 0.15 and 0.20, having accuracy not exceeding $0.5 \times 10^{-4}$.

**Solution:** We form Taylor series expansion using $y$ (0), $y'$ (0) = 1 and from the differential equation,

$$y'' + xy' + y = 0, \text{ we get } y''(0) = 0$$

$$y'''(x) = -xy'' - 2y' \qquad \therefore y'''(0) = -2$$

$$y^{(iv)}(x) = -xy''' - 3y'' \qquad \therefore y^{iv}(0) = 0$$

$$y^{v}(x) = -xy^{(iv)} - 4y''' \quad \therefore y^{v}(0) = 8$$

And in general, $y^{(2n)}(0) = 0, \; y^{(2n+1)}(0) = -2ny^{(2n-1)}(0) = (-1)^n 2^n.2!$

Thus, $\quad y(x) = x - \dfrac{x^3}{3} + \dfrac{x^5}{15} - ... + (-1)^n \dfrac{2^n n! x 2^{n+1}}{(2n+1)!} + ...$

This is an alternating series whose terms decrease. Using this, we form the solution for $y$ up to 0.2 as given below:

| $x$ | 0 | 0.05 | 0.10 | 0.15 | 0.20 |
|---|---|---|---|---|---|
| $y(x)$ | 0 | 0.0500 | 0.0997 | 0.1489 | 0.1973 |

**Check Your Progress**

1. How are Euler's method and Taylor's method related?
2. Why should we not use Euler's method for a larger range of x?

## 11.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. If we take $k = 1$, we get the Euler's method, $y_1 = y_0 + h f(x_0, y_0)$.
2. The method should not be used for a larger range of $x$ about $x_0$, since the propagated error grows as integration proceeds.

## 11.4 SUMMARY

- Euler's is a crude but simple method of solving a first order initial value problem:

$$\frac{dy}{dx} = f(x, y), \; y(x_0) = y_0$$

- The local error at any $x_k$, i.e., the truncation error of the Euler's method is given by,

$$e_k = y(x_{k+1}) - y_{k+1}$$

- Modified Euler's method can be used to compute the solution up to a desired accuracy by applying it in an iterative scheme as stated below.

Compute $y^{(k)}{}_{n+1} = y_n + h\, f(x_n, \; y_n)$

Compute $\; y_{n+1}^{(k+1)} = y_n + \dfrac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(k)})\right], \text{ for } k = 0, 1, 2, ...$

- Euler's method can be extended to compute approximate values $y_i$ and $z_i$ of $y(x_i)$ and $z(x_i)$ respectively given by,

  $y_{i+1} = y_i + h\, f(x_i, y_i, z_i)$

  $z_{i+1} = z_i + h\, g(x_i, y_i, z_i)$

## 11.5 KEY WORDS

- **Euler's method:** The Euler's method finds a sequence of values $\{y_k\}$ of $y$ for the sequence of values $\{x_k\}$ of $x$, step by step.

## 11.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

**Short-Answer Questions**

1. Define Euler's method.
2. Explain the Euler's method for a pair of differential equations.

**Long-Answer Questions**

1. Compute values of $y$ at $x = 0.02$, by Euler's method taking $h = 0.01$, given $y$ is the solution of the following initial value problem: $\frac{dy}{dx} = x^3 + y, y(0) = 1$.
2. Evaluate $y(0.02)$ by modified Euler's method, given $y' = x^2 + y, y(0) = 1$, correct to four decimal places.

## 11.7 FURTHER READINGS


Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

NOTES

# BLOCK - IV
# TAYLOR'S METHOD, R.K METHOD AND STABILITY ANALYSIS

# UNIT 12 TAYLOR'S METHOD

**Structure**



## 12.0 INTRODUCTION

In mathematics, the Taylor series of a function is an infinite sum of terms that are expressed in terms of the function's derivatives at a single point. For most common functions, the function and the sum of its Taylor series are equal near this point. Taylor's series are named after Brook Taylor who introduced them in 1715.

In this unit, you will study about the Taylor's method.

## 12.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the Taylor's method
- Explain the Taylor's series

## 12.2 TAYLOR'S METHOD

Consider the solution of the first order differential equation,

$$\frac{dy}{dx} = f(x, y) \text{ with } y(x_0) = y_0 \tag{12.1}$$

where $f(x, y)$ is sufficiently differentiable with respect to $x$ and $y$. The solution $y(x)$ of the problem can be expanded about the point $x_0$ by a Taylor series in the form,

$$y(x_0 + h) = y(x_0) + hy'(x_0) + \frac{h^2}{2!} y''(x_0) + \ldots + \frac{y^k(x_0)}{k!} h^k + \frac{h^{k+1}}{(k+1)!}(\xi) \tag{12.2}$$

The derivatives in the above expansion can be determined as follows,

$$y'(x_0) = f(x_0, y_0)$$
$$y''(x_0) = f_x(x_0, y_0) + f_y(x_0, y_0)\, y'(x_0)$$
$$y'''(x_0) = f_{xx}(x_0, y_0) + 2f_{xy}(x_0, y_0)\, y'(x_0) + f_{yy}(x_0, y_0)\{y'(x_0)\}^2 + f_y(x, y)\, y''(x_0)$$

where a suffix $x$ or $y$ denotes partial derivative with respect to $x$ or $y$.

Thus the value of $y_1 = y(x_0+h)$, can be computed by taking the Taylor series expansion shown above. Usually, because of difficulties in obtaining higher order derivatives, commonly a fourth order method is used. The solution at $x_2 = x_1+h$, can be found by evaluating the derivatives at $(x_1, y_1)$ and using the expansion; otherwise, writing $x_2 = x_0+2h$, we can use the same expansion. This process can be continued for determining $y_{n+1}$ with known values $x_n, y_n$.

***Note:*** If we take $k = 1$, we get the Euler's method, $y_1 = y_0+h f(x_0, y_0)$.

Thus, Euler's method is a particular case of Taylor series method.

**Example 1:** Form the Taylor series solution of the initial value problem, $\frac{dy}{dx} = xy + 1,\ y(0) = 1$ up to five terms and hence compute $y(0.1)$ and $y(0.2)$, correct to four decimal places.

**Solution:** We have, $y' = xy + 1,\ y(0) = 1$

Differentiating successively we get,

$$y''(x) = xy' + y,\ \therefore y''(0) = 1$$
$$y'''(x) = xy'' + 2y',\ \therefore y'''(0) = 2$$
$$y^{(iv)}(x) = xy''' + 3y'',\ \therefore y^{(iv)}(0) = 3$$
$$y^{(v)}(x) = xy^{(iv)} + 3y''',\ \therefore y^{(v)}(0) = 6$$

Hence, the Taylor series solution $y(x)$ is given by,

$$y(x) \approx y(0) + xy'(0) + \frac{x^2}{2} y''(0) + \frac{x^3}{3!} y'''(0) + \frac{x^4}{4!} y^{(iv)}(0) + \frac{x^5}{5!} y^{(v)}(0)$$
$$\approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6} \times 2 + \frac{x^4}{24} \times 3 + \frac{x^5}{120} \times 6 = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{8} + \frac{x^5}{20}$$
$$\therefore y(0.1) \approx 1 + 0.1 + \frac{0.01}{2} + \frac{0.001}{3} + \frac{0.0001}{8} + \frac{0.00001}{20} = 1.1053$$

Similarly, $y\ (0.2) \approx 1 + 0.2 + \frac{0.04}{2} + \frac{0.008}{3} + \frac{0.0016}{8} + \frac{0.00032}{20} = 1.04274$

**Example 2:** Find first two non-vanishing terms in the Taylor series solution of the initial value problem $y' = x^2 + y^2,\ y(0) = 0.$ Hence, compute $y(0.1), y(0.2), y(0.3)$ and comment on the accuracy of the solution.

**Solution:** We have, $y' = x^2 + y^2, \quad y \quad (0) = 0$

Differentiating successively we have,

**NOTES**

$$\begin{aligned}
y'' &= 2x + 2yy', & \therefore y''(0) &= 0 \\
y''' &= 2 + 2[yy'' + (y')^2, & y'''(0) &= 2 \\
y^{(iv)} &= 2(yy''' + 3y'y''), & \therefore y^{(iv)}(0) &= 0 \\
y^{(v)} &= 2[yy^{(iv)} + 4y'y''' + 3(y'')^2], & \therefore y^{(v)} &= 0 \\
y^{(vi)} &= 2[yy^{(v)} + 5y'y^{(iv)} + 10y''y'''], & \therefore y^{(vi)} &= 0 \\
y^{(vii)} &= 2[yy^{(vi)} + 6y'\,y^{(v)} + 15y''\,y^{(iv)} + 10(y''')^2] & \therefore y^{(vi)}(0) &= 80
\end{aligned}$$

The Taylor series up to two terms is $y(x) = \frac{x^3}{6} \times 2 + \frac{x^7}{7} \times \frac{80}{7!} = \frac{1}{3}x^3 + \frac{x^7}{63}$

**Example 3:** Given $x\,y' = x - y^2$, $y(2) = 1$, evaluate $y(2.1)$, $y(2.2)$ and $y(2.3)$ correct to four decimal places using Taylor series method.

**Solution:** Given $y' = x - y^2$, i.e., $y' = 1 - y^2/x$ and $y = 1$ for $x = 2$. To compute $y(2.1)$ by Taylor series method, we first find the derivatives of $y$ at $x = 2$.

$$y' = 1 - y^2/x \qquad \therefore \quad y'(2) = 1 - \frac{1}{2} = 0.5$$

$$xy'' + y' = 1 - 2yy'$$

$$2y''(2) + \frac{1}{2} = 1 - 2.\frac{1}{2} \qquad \therefore \quad y''(2) = \frac{1}{4} - \frac{2}{2} \times \frac{1}{2} = -0.25$$

$$xy''' + 2y'' = -2y'^2 - 2yy''$$

$$\therefore \quad 2y'''(2) + 2\left(-\frac{1}{4}\right) = -2\left(\frac{1}{2}\right)^2 - 2\left(-\frac{1}{4}\right)$$

Or, $$2y'''(2) = \frac{1}{2} \qquad \therefore \quad y'''(2) = \frac{1}{4} = 0.25$$

$$xy^{(iv)} + 3y''' = -4y'y'' - 2y'y'' - 2yy'''$$

$$2y''''(2) + 3 \times \frac{1}{4} = 6 \times \frac{1}{2} \times \left(\frac{1}{4}\right) - 2 \times \frac{1}{4}$$

$$y''''(2) = \left(\frac{3}{4} - \frac{3}{4} - \frac{1}{2}\right)\frac{1}{2} = -0.25$$

$$\begin{aligned}
y(2.1) &= y \quad (2) + 0.1 \quad y'(2) + \frac{(0.1)^2}{2}y''(2) + \frac{(0.1)^3}{3!}y'''(2) + \frac{(0.1)^4}{4!}y''''(2) \\
&= 1 + 0.1 \times 0.5 + \frac{0.01}{2} \times (-0.25) + \frac{0.001}{6} \times 0.25 + \frac{0.0001}{24} \times (-0.25) \\
&= 1 + 0.05 - 0.00125 + 0.00004 - 0.000001 \\
&= 1.0488
\end{aligned}$$

$$y(2.2)=1+0.2\times 0.5+\frac{0.04}{2}\times(-0.25)+\frac{0.008}{6}\times 0.25+\frac{0.0016}{24}\times(-0.5)$$

$$=1+0.1-0.005-0.00032-0.00003$$

$$=1.0954$$



$$y(2.3)=1+0.3\times 0.5+\frac{0.09}{2}\ (-0.25)+\frac{0.009}{2}\times 0.25+\frac{0.0081}{24}\times(0.5)$$

$$=1+0.15-0.01125+0.001125+0.000168$$

$$=1.005043$$

**Check Your Progress**

1. Explain the Taylor's method.
2. Dive the derivative of the Taylor's series.

## 12.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Consider the solution of the first order differential equation,

   $$\frac{dy}{dx}=f(x,y) \text{ with } y(x_0)=y_0$$

2. The derivatives of Taylor's series can be determined as follows:

$$y'(x_0)=f(x_0,y_0)$$

$$y''(x_0)=f_x(x_0,y_0)+f_y(x_0,y_0)\,y'(x_0)$$

$$y'''(x_0)=f_{xx}(x_0,\ y_0)+2f_{xy}(x_0,y_0)\,y'(x_0)+f_{yy}(x_0,\ y_0)\{y'(x_0)\}^2+f_y(x,y)y''(x_0)$$

## 12.4 SUMMARY

- The solution $y(x)$ of the problem can be expanded about the point $x_0$ by a Taylor series in the form,

  $$y(x_0+h)=y(x_0)+hy'(x_0)+\frac{h^2}{2!}y''(x_0)+...+\frac{y^k(x_0)}{k!}h^k+\frac{h^{k+1}}{(k+1)!}(\xi)$$

- because of difficulties in obtaining higher order derivatives, commonly a fourth order method is used.
- The solution at $x_2=x_1+h$, can be found by evaluating the derivatives at $(x_1, y_1)$ and using the expansion; otherwise, writing $x_2=x_0+2h$, we can use the same expansion.

NOTES

## 12.5 KEY WORDS

- **Taylor's series method:** If we take $k = 1$, we get the Euler's method, $y_1 = y_0 + h f(x_0, y_0)$.
  Thus, Euler's method is a particular case of Taylor series method.

## 12.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

**Short-Answer Questions**

1. What is Taylor's series?
2. Give one example of Taylor's method.

**Long-Answer Questions**

1. Discuss about the Taylor's method.
2. Compute the derivatives of Taylor's expansion.

## 12.7 FURTHER READINGS


Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

# UNIT 13 RUNGE KUTTA METHOD

NOTES

**Structure**



## 13.0 INTRODUCTION

In numerical analysis, the Runge–Kutta methods are a family of implicit and explicit iterative methods, which include the well-known routine called the Euler Method, used in temporal discretization for the approximate solutions of ordinary differential equations.

In this unit, you will study about the Runge-Kutta methods and Runge-Kutta methods for a pair of equations.

## 13.1 OBJECTIVES

After going through this unit, you will be able to:

- Analyse the Runge-Kutta methods
- Understand the Runge-Kutta methods for a pair of equations

## 13.2 RUNGE KUTTA METHOD

Runge-Kutta method can be of different orders. They are very useful when the method of Taylor series is not easy to apply because of the complexity of finding higher order derivatives. Runge-Kutta methods attempt to get better accuracy and at the same time obviate the need for computing higher order derivatives. These methods, however, require the evaluation of the first order derivatives at several off-step points.

Here we consider the derivation of Runge-Kutta method of order 2.

The solution of the $(n + 1)$th step is assumed in the form,

$$y_{n+1} = y_n + ak_1 + bk_2 \tag{13.1}$$

Where $k_1 = h\,f(x_n, y_n)$ and
$k_2 = h\,f(x_n + \alpha h, y_n + \beta k_1)$, for $n = 0, 1, 2,...$ (13.2)

The unknown parameters $a$, $b$, $\alpha$, and $\beta$ are determined by expanding in Taylor series and forming equations by equating coefficients of like powers of $h$. We have,

$$y_{n+1} = y(x_n + h) = y_n + h\,y'(x_n) + \frac{h^2}{2}y''(x_n) + \frac{h^3}{6}y'''(x_n) + 0(h^4)$$

$$= y_n + h\,f(x_n, y_n) + \frac{h^2}{2}[f_x + ff_y]_n + \frac{h^3}{6}[f_{xx} + 2ff_{yy} + f_{yy}f^2 + f_x f_y + f_y^2 f]_n + 0(h^4) \quad (13.3)$$

The subscript $n$ indicates that the functions within brackets are to be evaluated at $(x_n, y_n)$.

Again, expanding $k_2$ by Taylor series with two variables, we have

$$k_2 = h[f_n + ah \quad (f_x)_n + \beta k_1 (f_y)_n + \frac{\alpha^2\beta^2}{2} \quad (f_{xx})_n + \alpha\beta h k_1 \quad (f_{xy})_n + \frac{\beta^2 k_1^2}{2} \quad (f_{yy})_n + 0(h^3)] \quad (13.4)$$

Thus on substituting the expansion of $k_2$, we get from Equation (13.4)

$$y_{n+1} = y_n + (a+b)h \quad f_n + bh^2 \quad (\alpha f_x + \beta ff_y)_n + bh^3 \left(\frac{\alpha^2}{2} f_{xx} + \alpha\beta ff_{xx} + \frac{\beta^2}{2} f^2 f_{yy}\right) + 0(h^4)$$

On comparing with the expansion of $y_{n+1}$ and equating coefficients of $h$ and $h^2$ we get the relations,

$$a + b = 1, \quad b\alpha = b\beta = \frac{1}{2}$$

There are three equations for the determination of four unknown parameters. Thus, there are many solutions. However, usually a symmetric solution is taken by setting $a = b = \frac{1}{2}$, then $\alpha = \beta = 1$

Thus we can write a Runge-Kutta method of order 2 in the form,

$$y_{n+1} = y_n + \frac{h}{2} \quad [f(x_n, \quad y_n) + f(x_n + h, \quad y_n + h \quad f(x_n, \quad y_n))], \text{ for } n = 0, \ 1, \ 2,... \quad (13.5)$$

Proceeding as in second order method, Runge-Kutta method of order 4 can be formulated. Omitting the derivation, we give below the commonly used Runge-Kutta method of order 4.

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) + 0(h^5)$$

$$k_1 = h \ f(x_n, y_n)$$

$$k_2 = h\,f\left(x_n + \frac{h}{2}, \ y_n + \frac{k_1}{2}\right)$$

$$k_3 = h\,f\left(x_n + \frac{h}{2}, \ y_n + \frac{k_2}{2}\right)$$

$$k_4 = h\,f(x_n + h, \ y_n + k_3) \quad (13.6)$$

NOTES

Runge-Kutta method of order 4 requires the evaluation of the first order derivative $f(x, y)$, at four points. The method is self-starting. The error estimate with this method can be roughly given by,

$$|y(x_n) - y_n| \approx \frac{y_n^* - y_n}{15} \qquad (13.7)$$

where $y_n^*$ and $y_n$ are the approximate values computed with $\frac{h}{2}$ and $h$ respectively as step size and $y(x_n)$ is the exact solution.

***Note:*** In particular, for the special form of differential equation $y' = F(x)$, a function of $x$ alone, the Runge-Kutta method reduces to the Simpson's one-third formula of numerical integration from $x_n$ to $x_{n+1}$. Then,

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} F(x)dx$$

Or,
$$y_{n+1} = y_n + \frac{h}{6}[F(x_n) + 4F(x_n + \frac{h}{2}) + F(x_n + h)]$$

Runge-Kutta methods are widely used particularly for finding starting values at steps $x_1, x_2, x_3, \ldots$, since it does not require evaluation of higher order derivatives. It is also easy to implement the method in a computer program.

**Example 1:** Compute values of $y(0.1)$ and $y(0.2)$ by 4th order Runge-Kutta method, correct to five significant figures for the initial value problem,

$$\frac{dy}{dx} = x + y,\ y(0) = 1$$

**Solution:** We have $\frac{dy}{dx} = x + y,\ y(0) = 1$

$$\therefore f(x,\ y) = x + y, \quad h = 0.1, \quad x_0 = 0, \quad y_0 = 1$$

By Runge-Kutta method,

$$y(0.1) = y(0) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

where,
$$k_1 = h\, f(x_0,\ y_0) = 0.1 \times (0 + 1) = 0.1$$

$$k_2 = h\, f\left(x_0 + \frac{h}{2},\ y_0 + \frac{k_2}{2}\right) = 0.1 \times (0.05 + 1.05) = 0.11$$

$$k_3 = h\, f\left(x_0 + \frac{h}{2},\ y_0 + \frac{k_2}{2}\right) = 0.1 \times (0.05 + 1.055) = 0.1105$$

$$k_4 = h\, f(x_0 + h,\ y_0 + k_3) = 0.1 \times (0.1 + 1.1105) = 0.12105$$

where
$$\therefore\ y(0.1) = 1 + \frac{1}{6}\ [0.1 + 2 \times (0.11 + 0.1105 + 0.12105] = 1.130516$$

Thus, $x_1 = 0.1, \quad y_1 = 1.130516$

$$y(0.2) = y(0.1) + \frac{1}{6}\ (k_1 + 2k_2 + 2k_3 + k_4)$$

$$k_1 = h\ f(x_1,\ \ y_1) = 0.1 \times (0.1 + 1.11034) = 0.121034$$

$$k_2 = h\ f\left(x_1 + \frac{h}{2}, y_1 + \frac{k_1}{2}\right) = 0.1\ (0.15 + 1.17086) = 0.132086$$

$$k_3 = h\ f\left(x_1 + \frac{h}{2}, y_1 + \frac{k_2}{2}\right) = 0.1\ (0.15 + 1.17638) = 0.132638$$

$$k_4 = h\ f(x_1 + h, y_1 + k_3) = 0.1\ \ (0.2 + 1.24298) = 0.144298$$

$$y_2 = y\ (0.2) = 1.11034 + \frac{1}{6}\ \ [0.121034 + 2\ (0.132086 + 0.132638) + 0.144298] = 1.2428$$

**Example 2:** Use Runge-Kutta method of order 4 to evaluate $y\,(1.1)$ and $y\,(1.2)$, by taking step length $h = 0.1$ for the initial value problem,

$$\frac{dy}{dx} = x^2 + y^2,\ y(1) = 0$$

**Solution:** For the initial value problem,

$\frac{dy}{dx} = f(x,\ \ y),\ y\ (x_0) = y_0$, the Runge-Kutta method of order 4 is given as,

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

where

$$k_1 = h\ f(x_n,\ \ y_n)$$

$$k_2 = h\ f\left(x_n + \frac{h}{2},\ \ y_n + \frac{k_1}{2}\right)$$

$$k_3 = h\ f\left(x_n + \frac{h}{2},\ \ y_n + \frac{k^2}{2}\right)$$

$$k_4 = h\ f(x_n + h, y_n + k_3),\ \text{for } n = 0, 1, 2, ...$$

For the given problem, $f(x, y) = x^2 + y^2$, $x_0 = 1$, $y_0 = 0$, $h = 0.1$.

Thus,

$$k_1 = h\ f(x_0, y_0) = 0.1 \times (1^2 + 0^2) = 0.1$$

$$k_2 = h\ f\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = 0.1 \times [(1.05)^2 + (0.5)^2] = 0.13525$$

$$k_3 = h\ f\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = 0.1 \times [(1.05)^2 + (0.05525)^2] = 0.13555$$

$$k_4 = h\ f(x_0 + h, y_0 + k_3) = 0.1 \times [(1.1)^2 + (0.13555)^2] = 0.12283$$

$$\therefore y_1 = y_0 + \frac{1}{6}\ \ (k_1 + 2k_2 + 2k_3 + k_4)$$

$$= \frac{1}{6}(0.1 + 0.2705 + 0.2711 + 0.12283) = \frac{1}{6} \times 0.76443$$

$$= 0.127405$$

For $y$ (1.2):

$$k_1 = 0.1\,[(1.1)^2 + (0.11072)^2] = 0.12226$$
$$k_2 = 0.1\,[(1.15)^2 + (0.17183)^2] = 0.135203$$
$$k_3 = 0.1\,[(1.15)^2 + (0.17832)^2] = 0.135430$$
$$k_4 = 0.1\,[(1.2)^2 + (0.24615)^2] = 0.150059.$$

$$\therefore\ y_2 = y\,(1.2) = 0.11072 + \frac{1}{6}(0.12226 + 0.270406 + 0.270860 + 0.150069)$$
$$= 0.24631$$

NOTES

**Algorithm:** Solution of first order differential equation by Runge-Kutta method of order 2: $y' = f(x)$ with $y\,(x_0) = y_0$.

**Step 1:** Define $f(x, y)$

**Step 2:** Read $x_0, y_0, h, x_f$ [$h$ is step size, $x_f$ is final $x$]

**Step 3:** Repeat Steps 4 to 11 until $x_1 > x_f$

**Step 4:** Compute $k_1 = f(x_0, y_0)$

**Step 5:** Compute $y_1 = y_0 + hk_1$

**Step 6:** Compute $x_1 = x_0 + h$

**Step 7:** Compute $k_2 = f(x_1, y_1)$

**Step 8:** Compute $y_1 = y_0 + h \times (k_1 + k_2)/2$

**Step 9:** Write $x_1, y_1$

**Step 10:** Set $x_0 = x_1$

**Step 11:** Set $y_0 = y_1$

**Step 12:** Stop

**Algorithm:** Solution of $y_1 = f(x,\ y),\ \ y(x_0) = y_0$ by Runge-Kutta method of order 4.

**Step 1:** Define $f(x, y)$

**Step 2:** Read $x_0, y_0, h, x_f$

**Step 3:** Repeat Step 4 to Step 16 until $x_1 > x_f$

**Step 4:** Compute $k_1 = h\,f(x_0, y_0)$

**Step 5:** Compute $x = x_0 + \dfrac{h}{2}$

**Step 6:** Compute $y = y_0 + \dfrac{k_1}{2}$

**Step 7:** Compute $k_2 = h\,f(x, y)$

**Step 8:** Compute $y = y_0 + \dfrac{k_2}{2}$

**NOTES**

**Step 9:** Compute $k_3 = h f(x, y)$

**Step 10:** Compute $x_1 = x_0 + h$

**Step 11:** Compute $y = y_0 + k_3$

**Step 12:** Compute $k_4 = h f(x_1, y)$

**Step 13:** Compute $y_1 = y_0 + (k_1 + 2 (k_2 + k_3) + k_4)/6$

**Step 14:** Write $x_1, y_1$

**Step 15:** Set $x_0 = x_1$

**Step 16:** Set $y_0 = y_1$

**Step 17:** Stop

### Runge-Kutta Method for a Pair of Equations

Consider an initial value problem associated with a system of two first order ordinary differential equations in the form,

$$\frac{dy}{dx} = f(x, y, z), \frac{dz}{dx} = g(x, y, z)$$

with $y(x_0) = y_0$ and $z(x_0) = z_0$

The Runge-Kutta method of order 4 can be easily extended in the following form,

$$\begin{aligned} y_{i+1} &= y_i + \frac{1}{6} \ (k_1 + 2k_2 + 2k_3 + k_4) \\ z_{i+1} &= z_i + \frac{1}{6} \ (l_1 + 2l_2 + 2l_3 + l_4) \text{ for } i = 0, 1, 2,... \end{aligned} \tag{13.8}$$

Where

$$\begin{aligned} k_1 &= hf(x_i, y_i, z_i), & l_1 &= hg(x_i, y_i, z_i) \\ k_2 &= hf\left(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}, z_i + \frac{l_1}{2}\right), & l_2 &= hg\left(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}, z_1 + \frac{l_1}{2}\right) \\ k_3 &= hf\left(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}, z_i + \frac{l_2}{2}\right), & l_3 &= hg\left(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}, z_i + \frac{l_2}{2}\right) \\ k_4 &= hf(x_i + h, y_1 + k_3, z_i + l_3), & l_4 &= hf(x_i + h, y_i + k_3, z_i + l_3) \end{aligned}$$

$$y_i = y(x_i), z_i = z(x_i), i = 0, 1, 2,...$$

The solutions for $y(x)$ and $z(x)$ are determined at successive step points $x_1 = x_0 + h, x_2 = x_1 + h = x_0 + 2h,..., x_N = x_0 + Nh.$

### Runge-Kutta Method for a Second Order Differential Equation

Consider the initial value problem associated with a second order differential equation,

$$\frac{d^2 y}{dx^2} = g(x, y, y')$$

with $y(x_0) = y_0$ and $y'(x_0) = \alpha_0$

On substituting $z = y'$, the above problem is reduced to the problem,

NOTES

$$\frac{dy}{dx} = z, \frac{dz}{dx} = g(x, y, z)$$

with $y(x_0) = y_0$ and $z(x_0) = y'(x_0) = \alpha_0$

which is an initial value problem associated with a system of two first order differential equations. Thus we can write the Runge-Kutta method for a second order differential equation as,

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$
$$z_{i+1} = y'_{i+1} = z_i + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4) \text{ for } i = 0, 1, 2, ... \quad (13.9)$$

where $k_1 = h(z_i)$, $l_1 = hg(x_i, y_i, z_i)$

$$k_2 = h\left(z_i + \frac{l_1}{2}\right), \quad l_2 = hg\left(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}, z_i + \frac{l_1}{2}\right)$$
$$k_3 = h\left(z_i + \frac{l_2}{2}\right), \quad l_3 = hg\left(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}, z_i + \frac{l_2}{2}\right)$$
$$k_4 = h(z_i + l_3), \quad l_4 = hg(x_i + h, y_i + k_3, z_i + l_3)$$

**Check Your Progress**

1. When are Runge-Kutta methods applied?
2. Give the uses of Runge-Kutta method.

## 13.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Runge-Kutta methods are very useful when the method of Taylor series is not easy to apply because of the complexity of finding higher order derivatives.
2. Runge-Kutta methods are widely used particularly for finding starting values at steps $x_1, x_2, x_3, ...$, since it does not require evaluation of higher order derivatives. It is also easy to implement the method in a computer program.

## 13.4 SUMMARY

- Runge-Kutta methods attempt to get better accuracy and at the same time obviate the need for computing higher order derivatives.

NOTES

- The solution of the $(n + 1)$th step is assumed in the form,

  $$y_{n+1} = y_n + ak_1 + bk_2$$

  Where $k_1 = h\,f(x_n, y_n)$ and $k_2 = h\,f(x_n + \alpha h, y_n + \beta k_1)$, for $n = 0, 1, 2,...$
- Runge-Kutta method of order 4 requires the evaluation of the first order derivative $f(x, y)$, at four points. The method is self-starting.
- In particular, for the special form of differential equation $y' = F(x)$, a function of $x$ alone, the Runge-Kutta method reduces to the Simpson's one-third formula of numerical integration from $x_n$ to $x_{n+1}$.
- The Runge-Kutta method of order 4 can be easily extended in the following form,

  $$y_{i+1} = y_i + \frac{1}{6}\ (k_1 + 2k_2 + 2k_3 + k_4)$$

  $$z_{i+1} = z_i + \frac{1}{6}\ (l_1 + 2l_2 + 2l_3 + l_4) \text{ for } i = 0, 1, 2,...$$

## 13.5 KEY WORDS

- **Runge-Kutta Method** It can be of different orders. They are very useful when the method of Taylor series is not easy to apply because of the complexity of finding higher order derivatives.

## 13.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. What is the significance of Runge-Kutta methods of different orders?
2. Explain the Runge-Kutta method for a pair of equations.

### Long-Answer Questions

1. Using Runge-Kutta method of order 4, compute $y(0.1)$ for each of the following problems:

   (a) $\frac{dy}{dx} = x + y,\ y(0) = 1$

   (b) $\frac{dy}{dx} = x + y^2,\ y(0) = 1$
2. Compute solution of the following initial value problem by Runge-Kutta method of order 4 taking $h = 0.2$ upto $x = 1$; $y' = x - y$, $y(0) = 1.5$.

## 13.7 FURTHER READINGS



Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.

# UNIT 14 STABILITY ANALYSIS

NOTES

**Structure**



## 14.0 INTRODUCTION

In mathematics, **stability theory** addresses the stability of solutions of differential equations and of trajectories of dynamical systems under small perturbations of initial conditions. The heat equation, for example, is a stable partial differential equation because small perturbations of initial data lead to small variations in temperature at a later time as a result of the maximum principle. In partial differential equations one may measure the distances between functions using Lp norms or the sup norm, while in differential geometry one may measure the distance between spaces using the Gromov–Hausdorff distance.

In this unit, you will study about stability analysis.

## 14.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the basic concept of stability analysis
- Understand the use of stability concept in finding solutions

## 14.2 STABILITY ANALYSIS

In mathematics and statistics stability theory defines the stability of solutions of differential equations and of trajectories of dynamical systems under small perturbations of initial conditions. The heat equation, for example, is a stable partial differential equation because small perturbations of initial data lead to small variations in temperature at a later time as a result of the maximum principle. In partial differential equations one may measure the distances between functions using Lp norms or the sup norm, while in differential geometry one may measure the distance between spaces using the Gromov–Hausdorff distance.

Under favourable circumstances, the question may be reduced to a well-studied problem involving eigenvalues of matrices. A more general method involves Lyapunov functions. In practice, any one of a number of different stability criteria are applied.

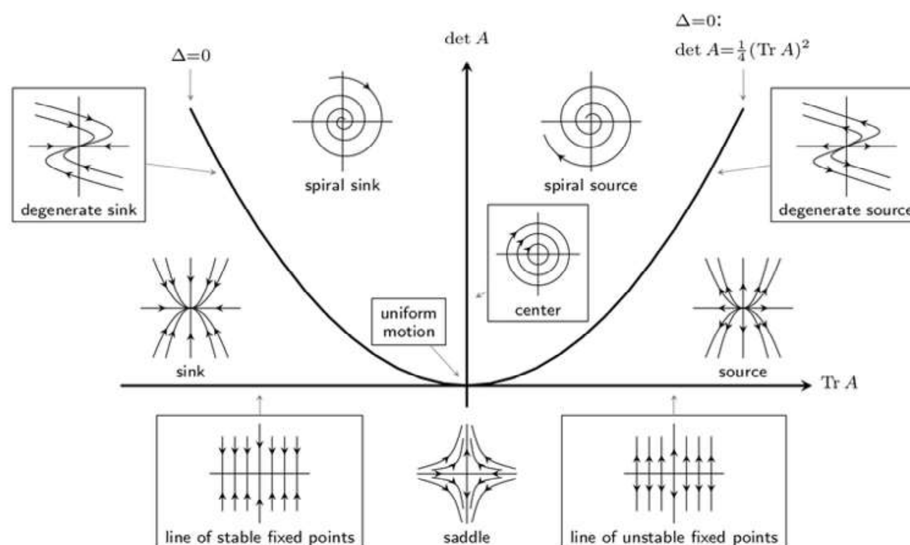
NOTES

### Overview in Dynamical Systems

Many parts of the qualitative theory of differential equations and dynamical systems deal with asymptotic properties of solutions and the trajectories—what happens with the system after a long period of time. The simplest kind of behaviour is exhibited by equilibrium points, or fixed points, and by periodic orbits. If a particular orbit is well understood, it is natural to ask next whether a small change in the initial condition will lead to similar behaviour. Stability theory helps us to find that whether a nearby orbit will indefinitely stay close to a given orbit ,or will it converge to a given orbit. In the former case, the orbit is called stable; in the latter case, it is called asymptotically stable and the given orbit is said to be attracting.

An equilibrium solution $f_e$ to an autonomous system of first order ordinary differential equations is called:

- Stable if for every (small) $\varepsilon > 0$, there exists a $\delta > 0$ ,such that every solution $f(t)$ having initial conditions within distance $\delta$ ,i.e. , $\| f(t_0) - , f_e \| < \delta$ of the equilibrium remains within distance $\varepsilon$, i.e. , for all $\| f(t) - f_e \| < \varepsilon$ for all $t \geq t_0$.
- Asymptotically stable if it is stable and, in addition, there exists $\delta_0 > 0$ , such that whenever $\| f(t_0) - f_e \| < \delta_0$ then $f(t) \rightarrow f_e$ as $t \rightarrow \infty$.



NOTES

Stability means that the trajectories do not change too much under small perturbations. The opposite situation, where a nearby orbit is getting repelled from the given orbit. In general, perturbing the initial state in some directions results in the trajectory asymptotically approaching the given one and in other directions to the trajectory getting away from it. There may also be directions for which the behaviour of the perturbed orbit is more complicated (neither converging nor escaping completely), and then stability theory does not give sufficient information about the dynamics.

In stability theory, the qualitative behaviour of an orbit under perturbations can be analysed using the linearization of the system near the orbit. In particular, at each equilibrium of a smooth dynamical system with an $n$-dimensional phase space, there is a certain $n \times n$ matrix $A$ whose eigenvalues characterize the behaviour of the nearby points (Hartman–Grobman theorem). More precisely, if all eigenvalues are negative real numbers or complex numbers with negative real parts then the point is a stable attracting fixed point, and the nearby points converge to it at an exponential rate, Lyapunov stability and exponential stability. If none of the eigenvalues are purely imaginary (or zero) then the attracting and repelling directions are related to the eigen-spaces of the matrix $A$ with eigenvalues whose real part is negative and, respectively, positive. Analogous statements are known for perturbations of more complicated orbits.

**Stability of Fixed Points**

The simplest kind of an orbit is a fixed point, or an equilibrium. If a mechanical system is in a stable equilibrium state then a small push will result in a localized motion, for example, small oscillations as in the case of a pendulum. In a system with damping, a stable equilibrium state is moreover asymptotically stable. On the other hand, for an unstable equilibrium, such as a ball resting on a top of a hill, certain small pushes will result in a motion with a large amplitude that may or may not converge to the original state. Stability of a nonlinear system can be deduced from the stability of its linearization.

**Maps:** Let $f$: R $\rightarrow$ R be a continuously differentiable function with a fixed point $a$, $f(a) = a$. Consider the dynamical system obtained by iterating the function $f$:

$$x_{n+1} = f(x_n), \quad n = 0, 1, 2, \ldots.$$

The fixed point $\alpha$ is stable if the absolute value of the derivative of $f$ at $\alpha$ is strictly less than 1, and unstable if it is strictly greater than 1. This is because near the point $a$, the function $f$ has a linear approximation with slope $f'(a)$:

$$f(x) \approx f(a) + f'(a)(x - a).$$

NOTES

Thus

$$x_{n+1} - x_n = f(x_n) - x_n \simeq f(a) + f'(a)(x_n - a) - x_n = a + f'(a)(x_n - a) - x_n$$
$$= (f'(a) - 1)(x_n - a) \to \frac{x_{n+1} - x_n}{x_n - a} = f'(a) - 1$$

which means that the derivative measures the rate at which the successive iterates approach the fixed point $a$ or diverge from it. If the derivative at $a$ is exactly 1 or “1, then more information is needed in order to decide stability.

There is an analogous criterion for a continuously differentiable map $f$: $R^n \to R^n$ with a fixed point $a$, expressed in terms of its Jacobian matrix at $a$, $J_a(f)$. If all eigenvalues of $J$ are real or complex numbers with absolute value strictly less than 1 then $a$ is a stable fixed point; if at least one of them has absolute value strictly greater than 1 then $a$ is unstable. Just as for $n$=1, the case of the largest absolute value being 1 needs to be investigated further — the Jacobian matrix test is inconclusive. The same criterion holds more generally for diffeomorphisms of a smooth manifold.

### Linear Autonomous Systems

The stability of fixed points of a system of constant coefficient linear differential equations of first order can be analysed using the eigenvalues of the corresponding matrix.

An autonomous system

$x' = Ax$,

where $x(t) \in R^n$ and $A$ is an $n \times n$ matrix with real entries, has a constant solution

$x(t) = 0$.

(In a different language, the origin $0 \in R^n$ is an equilibrium point of the corresponding dynamical system.) This solution is asymptotically stable as $t \to \infty$ (“in the future”) iff for all eigenvalues $\lambda$ of $A$, $Re(\lambda) < 0$. Similarly, it is asymptotically stable as $t \to -\infty$ (“in the past”) iff for all eigenvalues $\lambda$ of $A$, $Re(\lambda) > 0$. If there exists an eigenvalue $\lambda$ of $A$ with $Re(\lambda) > 0$ then the solution is unstable for $t$ ’! “.

The stability of the origin for a linear system can be determined by the Routh–Hurwitz stability criterion. The eigenvalues of a matrix are the roots of its characteristic polynomial. A polynomial in one variable with real coefficients is called a Hurwitz polynomial if the real parts of all roots are strictly negative. The Routh–Hurwitz theorem implies a characterization of Hurwitz polynomials by means of an algorithm that avoids computing the roots.

### Non-Linear Autonomous Systems

Asymptotic stability of fixed points of a non-linear system can be demonstrated using the Hartman–Grobman theorem.

NOTES

Suppose that $v$ is a $C^1$-vector field in $R^n$ which vanishes at a point $p$, $v(p) = 0$. Then the corresponding autonomous system

$x' = v(x)$

has a constant solution

$x(t) = p.$

Let $J_p(v)$ be the $n \times n$ Jacobian matrix of the vector field $v$ at the point $p$. If all eigenvalues of $J$ have strictly negative real part then the solution is asymptotically stable. This condition can be tested using the Routh–Hurwitz criterion.

**Check Your Progress**

1. How can linear differential equations of first order be analysed?
2. Define Hurwitz polynomial.
3. How can asymptotic stability of fixed points be demonstrated?

## 14.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The stability of fixed points of a system of constant coefficient linear differential equations of first order can be analysed using the eigenvalues of the corresponding matrix.
2. A polynomial in one variable with real coefficients is called a Hurwitz polynomial if the real parts of all roots are strictly negative.
3. Asymptotic stability of fixed points of a non-linear system can be demonstrated using the Hartman–Grobman theorem.

## 14.4 SUMMARY

- The simplest kind of behaviour is exhibited by equilibrium points, or fixed points, and by periodic orbits.
- Stability theory helps us to find that whether a nearby orbit will indefinitely stay close to a given orbit ,or will it converge to a given orbit. In the former case, the orbit is called stable; in the latter case, it is called asymptotically stable and the given orbit is said to be attracting.
- Stability means that the trajectories do not change too much under small perturbations.
- In stability theory, the qualitative behaviour of an orbit under perturbations can be analysed using the linearization of the system near the orbit.

NOTES

- If all eigenvalues are negative real numbers or complex numbers with negative real parts then the point is a stable attracting fixed point, and the nearby points converge to it at an exponential rate, Lyapunov stability and exponential stability.
- The simplest kind of an orbit is a fixed point, or an equilibrium.
- Stability of a nonlinear system can be deduced from the stability of its linearization.
- The fixed point $\alpha$ is stable if the absolute value of the derivative of $f$ at $\alpha$ is strictly less than 1, and unstable if it is strictly greater than 1.
- There is an analogous criterion for a continuously differentiable map $f$: $R^n \to R^n$ with a fixed point $a$, expressed in terms of its Jacobian matrix at $a$, $J_a(f)$.
- The stability of the origin for a linear system can be determined by the Routh–Hurwitz stability criterion.
- The Routh–Hurwitz theorem implies a characterization of Hurwitz polynomials by means of an algorithm that avoids computing the roots.
- If all eigenvalues of $J$ have strictly negative real part then the solution is asymptotically stable.

## 14.5 KEY WORDS

- **Maps:** Let $f$: $R \to R$ be a continuously differentiable function with a fixed point $a$, $f(a) = a$. Consider the dynamical system obtained by iterating the function $f$:

  $x_{n+1} = f(x_n), \quad n = 0, 1, 2, \ldots.$

## 14.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

### Short-Answer Questions

1. Define stability.
2. Elaborate on non-linear autonomous systems.

### Long-Answer Questions

1. Explain the stability theory.
2. Give details on linear autonomous systems.

## 14.7 FURTHER READINGS

NOTES

Jain, M. K., S. R. K. Iyengar and R. K. Jain. 2007. *Numerical Methods for Scientific and Engineering Computation*. New Delhi: New Age International (P) Limited.

Atkinson, Kendall E. 1989. *An Introduction to Numerical Analysis*, 2nd Edition. US: John Wiley & Sons.

Jain, M. K. 1983. *Numerical Solution of Differential Equations*. New Delhi: New Age International (P) Limited.

Conte, Samuel D. and Carl de Boor. 1980. *Elementary Numerical Analysis: An Algorithmic Approach*. New York: McGraw-Hill.

Skeel, Robert. D and Jerry B. Keiper. 1993. *Elementary Numerical Computing with Mathematica*. New York: McGraw-Hill.

Balaguruswamy, E. 1999. *Numerical Methods*. New Delhi: Tata McGraw-Hill.

Datta, N. 2007. *Computer Oriented Numerical Methods*. New Delhi: Vikas Publishing House Pvt. Ltd.